Target enrichment of long open reading frames and ultraconserved elements to link microevolution and macroevolution in non-model organisms

Claudia M Ortiz-Sepulveda, Mathieu Genete, Christelle Blassiau, Cécile Godé, Christian Albrecht, Xavier Vekemans, Bert Van Bocxlaer

Received: 14 November 2021 | Revised: 9 October 2022 | Accepted: 19 October 2022

DOI: 10.1111/1755-0998.13735

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Target enrichment of long open reading frames and ultraconserved elements to link microevolution and macroevolution in non-model organisms

Claudia M. Ortiz-Sepulveda[1] | Mathieu Genete[1] | Christelle Blassiau[1] | Cécile Godé[1] | Christian Albrecht[2,3] | Xavier Vekemans[1] | Bert Van Bocxlaer[1]

[1]CNRS, Univ. Lille, UMR 8198 – Evo-Eco-Paleo, F-59000 Lille, France

[2]Department of Animal Ecology and Systematics, Justus Liebig University, D-35392 Giessen, Germany

[3]Department of Biology, Mbarara University of Science and Technology, Mbarara, Uganda

Correspondence
Bert Van Bocxlaer and Claudia M. Ortiz-Sepulveda, CNRS, University of Lille, UMR 8198 – Evo-Eco-Paleo, F-59000 Lille, France.
Emails: bert.van-bocxlaer@univ-lille.fr; claudia.ortiz-sepulveda@univ-lille.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-JCJC-EVOLINK, ANR-17-CE02-0015, ANR-10-EQPX-07-01 and ANR-10-LABX-46; FEDER-ERC-EVORAD; French Ministère de l'Enseignement Supérieur et de la Recherche, Grant/Award Number: CPER_CLIMIBIO; European Fund for Regional Development (FEDER), Grant/Award Number: FEDER-ERC-EVORAD and CPER_CLIMIBIO; Hauts-de-France (HdF)

Handling Editor: Andrew DeWoody

Abstract

Despite the increasing accessibility of high-throughput sequencing, obtaining high-quality genomic data on non-model organisms without proximate well-assembled and annotated genomes remains challenging. Here, we describe a workflow that takes advantage of distant genomic resources and ingroup transcriptomes to select and jointly enrich long open reading frames (ORFs) and ultraconserved elements (UCEs) from genomic samples for integrative studies of microevolutionary and macroevolutionary dynamics. This workflow is applied to samples of the African unionid bivalve tribe Coelaturini (Parreysiinae) at basin and continent-wide scales. Our results indicate that ORFs are efficiently captured without prior identification of intron-exon boundaries. The enrichment of UCEs was less successful, but nevertheless produced substantial data sets. Exploratory continent-wide phylogenetic analyses with ORF supercontigs (>515,000 parsimony informative sites) resulted in a fully resolved phylogeny, the backbone of which was also retrieved with UCEs (>11,000 informative sites). Variant calling on ORFs and UCEs of Coelaturini from the Malawi Basin produced ~2000 SNPs per population pair. Estimates of nucleotide diversity and population differentiation were similar for ORFs and UCEs. They were low compared to previous estimates in molluscs, but comparable to those in recently diversifying Malawi cichlids and other taxa at an early stage of speciation. Skimming off-target sequence data from the same enriched libraries of Coelaturini from the Malawi Basin, we reconstructed the maternally-inherited mitogenome, which displays the gene order inferred for the most recent common ancestor of Unionidae. Overall, our workflow and results provide exciting perspectives for integrative genomic studies of microevolutionary and macroevolutionary dynamics in non-model organisms.

KEYWORDS
African freshwater molluscs, gene capture, genome skimming, phylogenetics, population genetics, transcriptomics (RNA-seq)

Claudia M. Ortiz-Sepulveda and Bert Van Bocxlaer contributed equally to the study.

1 | INTRODUCTION

Our planet's current biodiversity is intricately complex, dynamic and heterogenous, but the evolutionary history of most taxa is poorly documented. Such lack of knowledge may hamper conservation strategies, and, given current anthropogenic pressures on ecosystems (Barnosky et al., 2012; Dirzo et al., 2014), it may contribute to the irretrievable loss of poorly known biodiversity. Genomic data at the inter- and/or intraspecific level are essential to assess or infer ecological and evolutionary properties, but despite the continuous development of increasingly versatile genomic methods for non-model organisms (Romiguier et al., 2014), the differential rates at which genomic resources are acquired along various branches of the tree of life contribute to this heterogeneity (e.g., Gayral et al., 2013). Currently, detailed insights into the demographic history, the genome-wide development of reproductive isolation, the development of phenotypic traits and mechanisms of adaptation exist for a restricted number of model organisms with superior genomic resources (e.g., Cooney et al., 2017; Ronco et al., 2020; Van Belleghem et al., 2017).

A critical challenge for model and non-model organisms alike remains understanding the interplay of microevolutionary and macroevolutionary drivers and dynamics of diversification (Erwin, 2000; Reznick & Ricklefs, 2009), because both have traditionally been studied with different approaches and at different timescales. Microevolutionary studies typically consist of in-depth analyses of a small number of species based on large intraspecific samples, with limited opportunities for generalization across taxa. Macroevolutionary studies usually construe from comparative analyses on a restricted set of representatives for each of a large number of extant and/or fossil species, limiting insight into mechanisms operating at the level of individual species. Theoretical and empirical studies at both levels have indicated the need of a large number of orthologous loci to document phylogenetic relatedness, genetic diversity and population history (Dutoit et al., 2017; Helyar et al., 2011; Leaché & Rannala, 2011; Wortley et al., 2005). Rapid advances in high-throughput sequencing methods and genomic data analysis now allow to develop large multilocus data sets on model and non-model organisms alike, creating novel perspectives for the integration of micro- and macroevolutionary dynamics.

A multitude of strategies exist to obtain molecular data sets at a variety of taxonomic levels, enabling the development of genomic sampling schemes that could tackle questions at macroevolutionary and microevolutionary scales simultaneously. For non-model organisms the majority of approaches consist of reduced representation sequencing, where a subset of orthologous markers of the nuclear genome across taxa or individuals is obtained, for example with RAD-seq (Miller et al., 2007), transcriptomics (RNA-seq; Gayral et al., 2013), or by sequencing libraries after targeted sequence capture/enrichment (Hyb-seq). The latter strategy includes anchored hybrid enrichment (Lemmon et al., 2012), the sequencing of ultraconserved elements (UCEs; Faircloth et al., 2012), sequence capture using PCR-generated probes (Peñalba et al., 2014), transcriptome-based exon capture (Bi et al., 2012), exome capture with cDNA probes developed from expressed mRNA (Eec-seq; Puritz & Lotterhos, 2018), or if genomic resources exist, by using conserved noncoding elements (CNEs; Vavouri et al., 2007), including conserved nonexonic elements (CNEEs; Edwards et al., 2017).

Here, we develop a methodological framework combining several of the abovementioned approaches to enrich a set of loci that enables phylogenetic and population genetic studies in taxa with very limited genomic resources. Several motivations drive this effort. First, as comprehensive phylogenetic and population genetic studies require large sample sizes, we require a strategy that is scalable to 100 or 1000s of individuals without massive inflation of sequencing costs. Second, micro- and macroevolutionary studies each impose specific constraints, for example, related to orthology, the identification of coding versus noncoding regions and within coding regions of synonymous versus non-synonymous sites, so that the advantages and disadvantages of an integrative strategy are to be evaluated. Third, given the aim of integrative studies, it is desirable to select targets with a high information content. Fourth, in the absence of a well-assembled reference genome for the focal taxa or their close relatives, it remains difficult to leverage many of the abovementioned reduced representation sequencing methods. Specifically, we propose a strategy based on target enrichment of entire open reading frames (ORFs) of genes, which have been selected from ingroup-specific transcriptome sequencing, supplemented with more universal UCE targets that were identified from comparisons among distant genomes. The ORF of a gene is a stretch of DNA sequence, in the correct register, between start and stop codons that encodes a protein for translation, i.e. the coding sequence. Because ORFs are usually clustered in gene-rich regions within animal genomes (Osbourn & Field, 2009; Sproul et al., 2005), we include UCEs to increase the evenness at which the genome is sampled. Additionally, the integration of multiple types of markers has been suggested to enhance opportunities to resolve phylogenetic conflict (Chan et al., 2020; Hutter et al., 2019; Reddy et al., 2017). Both ORFs and UCEs are useful markers for organisms with no or limited genomic resources (Faircloth et al., 2012; Portik et al., 2016), as they enable the reconstruction of phylogenetic relationships across clades of varying age and taxonomic scale (Bi et al., 2012; Bragg et al., 2016; Faircloth et al., 2012; Harvey et al., 2016; Hugall et al., 2016; Lemmon et al., 2012; Teasdale et al., 2016), and they allow the detection of SNPs for population-level analyses (De Wit & Palumbi, 2012; Harvey et al., 2016; Schunter et al., 2014). A novelty of our approach is to focus on the entire ORFs of genes, which enhances the information content per marker and allows more rigorous assessment of genetic diversity at the population level, for example, through more accurate characterization of synonymous versus non-synonymous genetic diversity and demographic history (Gayral et al., 2013), including examinations of the speciation continuum (Roux et al., 2016). Inclusion of multiple exons per gene also provides access to additional intronic/intergenic flanking regions (compared to when a single exon is used), which may contain substantial

phylogenetic information, especially at shallow taxonomic levels (Breinholt et al., 2018).

Given the abovementioned requirements, we avoided RAD-seq, which produces short, blind markers for which alignment and orthology assessment may be challenging, especially for more divergent species. Additionally, it may be complicated to repeat, compare and combine various RAD-seq data sets, because the overlap of orthologous markers between data sets may be limited (Harvey et al., 2016). A further constraint of using RAD-seq data in the absence of a reference genome is that reads are to be analysed without alignment to an a priori determined sequence, so that the genetic diversity in RAD-seq data sets can be influenced by the level of natural heterozygosity in the studied samples and the parameters used for orthology assessment (Harvey et al., 2016). Orthology assessments may also be challenging for Eec-seq, which produces comparatively short markers that are not optimal for reconstructing genealogy at various levels of phylogenetic divergence. Subjecting all samples to transcriptome sequencing was not feasible because it does not allow leveraging historical, ethanol-preserved collections nor pooling as many samples per sequencing run. Consequently, it would result in decreased species/specimen representation and/or inflated costs. Additionally, levels of heterozygosity may be underestimated if only one of both alleles is expressed. Sequence capture approaches suffer less from the abovementioned issues, but face two important challenges to select targets: (1) the qualification of orthology, as only single-copy markers that are orthologous across all taxa under study are phylogenetically informative (Teasdale et al., 2016), and (2) the need to identify intron-exon boundaries to select exome targets (Karin et al., 2019; Portik et al., 2016). Both of these challenges typically require a reference genome (Bi et al., 2012; Bragg et al., 2016; Portik et al., 2016). In invertebrates, especially the molluscs we are concerned with here, orthology assessments are usually undertaken with very distant genomes due to the paucity of well-assembled genomes, for example, divergence >400 Ma for *Lottia* versus Eupulmonata in gastropods and *Bathymodiolus* versus Unionidae in bivalves (Combosh et al., 2017; Pfeiffer et al., 2019; Sun et al., 2019; Teasdale et al., 2016). Such ancient divergences imply that orthologue assessments for the reference may differ substantially from those for ingroup taxa. Here, we relax the need of well-assembled reference genomes by assessing orthology from existing genomic databases and representative ingroup transcriptomes. Additionally, by focusing on entire ORFs as functional biological units, instead of individual exons, we do not require to establish intron-exon boundaries prior to target enrichment. Whereas our proposed strategy enhances versatility, various issues could complicate target enrichment of entire ORFs, notably their subdivision in multiple exons. If exons are regularly shorter than the probe length, many probes will be tiled over exon boundaries within ORFs, which could drastically reduce the enrichment efficiency in genomic libraries. Evaluation of various Metazoan genomes indicated that genes consist on average of several short exons (number: 8.20 ± 1.90; length: 196 ± 69 bp; mean ± SD) that are separated from one another by much longer introns (length: 3079 ± 2063 bp) (Zhu et al., 2009). The number and lengths of exons in transcripts, the length of probes, the level of divergence between probes and targets and the length distribution of genomic library fragments are all important factors that could influence the success of our proposed strategy.

In light of the abovementioned considerations, we here describe an enrichment strategy for integrative studies of microevolution and macroevolution in the Afrotropical freshwater bivalve tribe Coelaturini (Parreysiinae; Unionidae) that can be readily expanded to other non-model organisms. We present a new approach to select orthologous single-copy genes from ingroup transcriptome assemblies, partly based on manual data curation (see Teasdale et al., 2016) and a strategy to successfully enrich their entire ORFs in genomic libraries. Additionally, we developed (to our knowledge) the first set of UCEs for both bivalves and gastropods (but see Moles & Giribet, 2021 for a gastropod-specific UCE set). We evaluate the performance of target enrichment for these heterogenous targets, and analyse the obtained data sets to illustrate their value for phylogenetics and population genetics. Finally, we built a custom, versatile pipeline to skim the raw sequencing data and to evaluate the possibility of recuperating off-target mitochondrial sequences, on which previous Sanger-sequencing studies of Unionidae (Lopes-Lima, Froufe, et al., 2017; Ortiz-Sepulveda et al., 2020; Whelan et al., 2011), and bivalves in general (Combosh et al., 2017), have relied heavily.

2 | MATERIALS AND METHODS

The workflow described below is summarized in Figure 1.

2.1 | RNA extraction and sequencing

Foot and mantle tissues of Coelaturini specimens from Malawi, Uganda and Zambia were stabilized in RNA-later for subsequent RNA extraction. Extractions were performed with the NucleoSpin RNA Plus kit of Machery-Nagel, either according to the protocol of the manufacturer, or by adding Proteinase K to the lysis buffer. Tissues were disrupted with magnetic beads (matrix D, 3 × 30 s at a speed of 6 ms, MP). The quantity and quality of mRNA was quantified with high-sensitivity Qubit fluorometry (Life Technologies Inc.) and with an Agilent BioAnalyser. We selected 12 samples from several major clades of Coelaturini based on the phylogeny of Ortiz-Sepulveda et al. (2020) for sequencing: Six specimens of *Coelatura hypsiprymna* and *C. nyassaensis* from the Malawi Basin; two of *Grandidieria burtoni* from Lake Tanganyika; two of *Coelatura hauttecoeuri* and two of *Nitia acuminata* from Lake Victoria (Table S1). RNA extraction products were purified and poly-A enriched for cDNA strand-specific, paired-end library preparation with the TruSeq RNA sample preparation kit version 2 (Illumina). The resulting libraries were sequenced on an Illumina HiSeq 3000 platform version 3.0 (2 × 150 bp). Purification, library preparation and sequencing were outsourced to UMR AGAP of the INRA and the University of Montpellier.

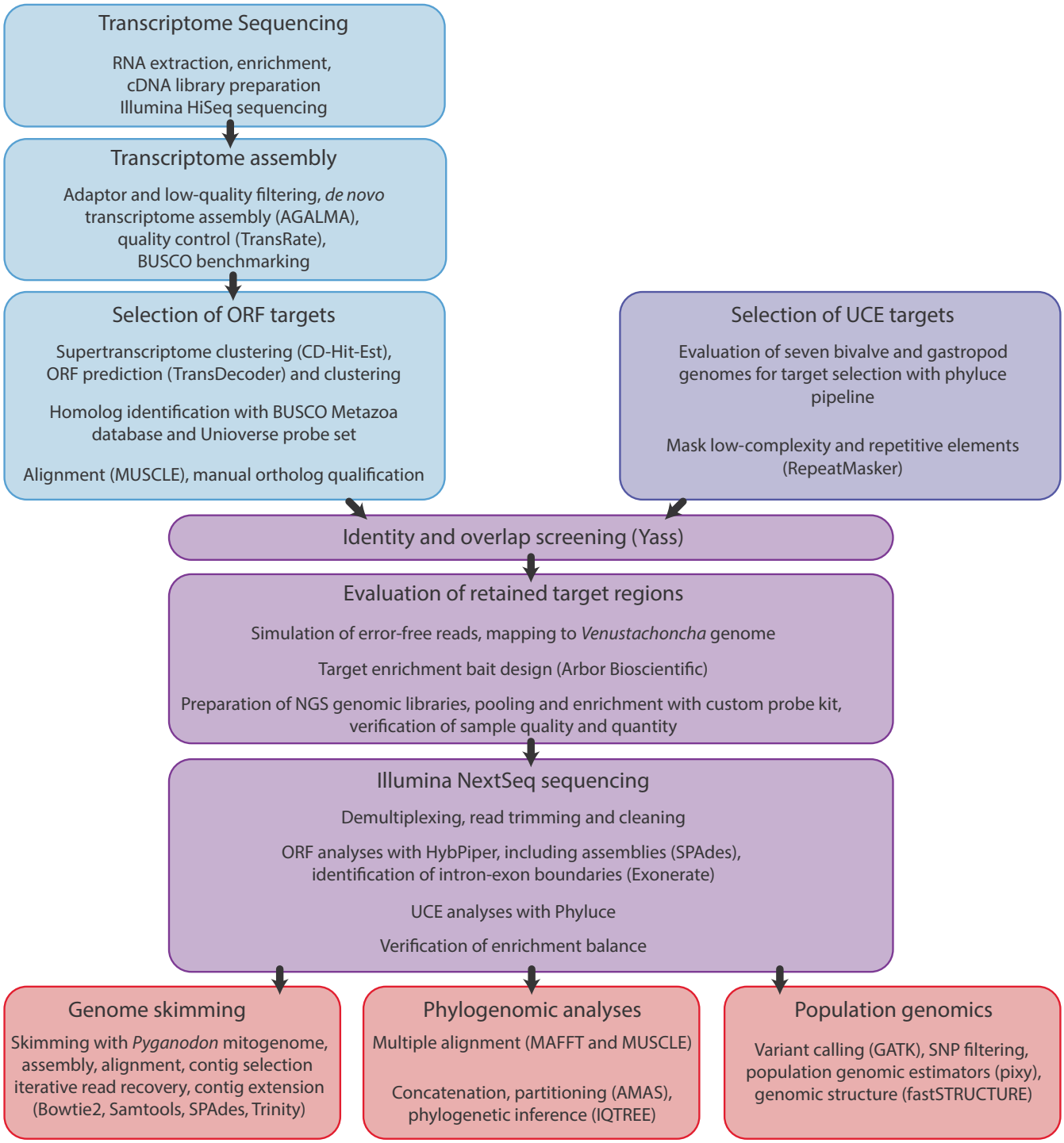

FIGURE 1 Schematic representation of our workflow indicating the main steps of the analytical pipeline. Our four main phases are represented in different colours; that is, transcriptomic steps for the selection of open reading frames (ORFs, blue), comparative genomics for the selection of ultraconserved elements (UCEs, dark purple), probe design and the sequencing of genomic libraries (light purple), analyses after the processing of sequencing data (red).

2.2 | Transcriptome assembly

After transcriptome sequencing we removed adaptor sequences and low-quality reads in multiple steps using TrimGalore! version 0.6.6 (Krueger, 2019), while iteratively examining the quantity and quality of reads with FastQC version 0.11.8 (Andrews, 2010). The cleaned reads were subjected to de novo transcriptome assembly with Agalma version 2.0.0 (Dunn et al., 2013), which includes several filtering procedures and ribosomal RNA removal followed by assembly with Trinity version 2.8.5 (Grabherr et al., 2011). We screened against the NCBI UniVec database to identify vector contaminants and rRNA transcripts. Subsequently, we performed

a reference-free assessment of the quality and completeness of our de novo assemblies with TRANSRATE version 1.0.3 (Smith-Unna et al., 2016) combined with BUSCO version 3.0.1 (Simão et al., 2015).

2.3 | Selection of ORF targets

To identify target ORFs for orthology assessment, we first clustered the contigs of our 12 AGALMA transcriptomes into a supertranscriptome using CD-HIT-EST version 4.8.1 (Li & Godzik, 2006) with a 95% similarity threshold, retaining the longest variant per cluster. We predicted ORFs for the supertranscriptome with TRANSDECODER version 5.5.0 (Haas & Papanicolaou, 2018) in two steps. First, we identified all likely coding regions from which we retained long ORFs, after which we selected the best-supported ORF per transcript. These ORF predictions were clustered again with CD-HIT-EST at 95% similarity to create homologous clusters. These homologous ORF clusters were expected to still contain a high level of redundancy because of fragmentation, frameshifts, mis-indexing, mis-assembly or the potential existence of isoforms or paralogues. We evaluated orthology in two steps. First, we mapped our ORF clusters against the BUSCO METAZOA_ODB9 database (Waterhouse et al., 2017) and against the UNIOVERSE probe set (Pfeiffer et al., 2019) to identify targets that have high chances of being single-copy and orthologous (=candidate genes). This approach does not allow to detect unexpressed homologues or weakly divergent paralogues, so additional follow-up verifications are performed during probe design (see below). We extracted all complete BUSCO hits, which consist of single- and multicopy hits. For ORFs with single-copy BUSCO hits the likelihood of multiple gene copies is small, given the gene is included in the single-copy orthologous database of BUSCO, and given the lack of expression variants with <95% similarity after clustering 12 ingroup transcriptomes. Multicopy BUSCO hits imply that at least two ORF homologous clusters map to a BUSCO gene and these were manually curated to distinguish among the following scenarios: (1) multiple copies may indicate the existence of distant homologues or paralogues, (2) assembly errors may have resulted in the creation of multiple contigs for the same gene, (3) various isoforms may exist of a single expressed RNA fragment. All ORFs that map to an individual BUSCO were merged into a .fasta file after which contigs were aligned with MUSCLE version 3.8.1551 (Edgar, 2004) and the nucleotide and protein sequences of homologues were visually inspected in SEAVIEW version 5 (Galtier et al., 1996; Gouy et al., 2010). If the presence of paralogues was suspected in this evaluation, we rejected all ORFs for the respective BUSCO hit. If different splicing or minor difficulties in assembly were suspected, the longest ORF was retained.

A similar approach was used to evaluate the 811 loci from the UNIOVERSE probe set, which was developed for anchored hybrid enrichment using the distant *Bathymodiolus platifrons* genome (Pfeiffer et al., 2019). This probe set contains 173,707 nucleotides, on average 214 per locus. To avoid overlap with targets that were already retained from the abovementioned BUSCO comparisons, we first screened the UNIOVERSE loci against the BUSCO METAZOA_ODB9 database and against our already retained ORFs with YASS version 1.15 (Noé & Kucherov, 2005). In total 109 UNIOVERSE loci were accounted for by these verifications, leaving 702 loci to be examined. Our remaining ORF clusters produced hits on all 702 remaining UNIOVERSE loci, and all hits for the same UNIOVERSE locus were compiled and aligned. When several UNIOVERSE loci mapped onto the same ORF, we also performed alignments including all concerned UNIOVERSE loci and all associated ORFs in SEAVIEW. The subsequent evaluation of candidate genes followed the criteria indicated above for our BUSCO evaluation. Finally, we mapped all retained ORF targets and their subregions to one another with YASS to examine the percentage of sequence similarity over a certain alignment length. ORFs were removed if alignment lengths and similarities were judged to potentially interfere with target enrichment.

2.4 | Selection of ultraconserved elements

We expanded our genomic sampling by constructing the (to our knowledge) first UCE probe set for gastropods and bivalves. We used seven published bivalve and gastropod genomes (Table S2), with the ampullariid gastropod *P. canaliculata* (Sun et al., 2019) as reference, to detect shared UCEs using PHYLUCE version 1.6.8 (Faircloth, 2016, 2017; Faircloth et al., 2012). Upon obtaining candidate UCEs, we masked repetitive elements and low complexity regions with REPEATMASKER version 4.0.9 (Smit et al., 2019), and merged the results for each genome in a table with commands of PHYLUCE. We retained UCEs with a minimum length of 100 bp (the length of our probes) that were retrieved from seven, six or five of the genomes, respectively. Subsequently, candidate UCEs were mapped onto our previously retained ORFs with YASS to examine potential overlaps and if so, the involved UCEs were discarded. We also mapped UCEs onto one another to avoid the inclusion of multiple UCEs with similar sequences.

2.5 | Evaluation of target regions

If exons are on average small, many probes may span exon boundaries within ORFs and may be inefficient for subsequent target enrichment with genomic DNA. We examined this issue using the *Venustaconcha ellipsiformis* genome (Renaut et al., 2018), the only genome for Unionidae available at the time. We generated error-free reads for our target regions (ORFs and UCEs) with ART (ART_ILLUMINA) version 2.5.8 (Huang et al., 2012). Each read had a length of 100 nt and reads were tilled to cover each base at 4×, resulting in a total of 84,848 reads. These reads were mapped to the *Venustaconcha* genome with STAMPY version 1.0.32 (Lunter & Goodson, 2011) and mapping statistics were examined with SAMTOOLS version 1.10 (Li et al., 2009). Subsequently, we produced a .bed file with functions of BEDTOOLS version 2.29 (Quinlan & Hall, 2010), and we used IGV version 2.6.3 (Robinson et al., 2011) to visualize hits for a subset of ORFs and UCEs.

2.6 | Target enrichment bait design

Target regions were submitted to Arbor Bioscientific for development of a custom MyBaits kit. Probes were 100 nt in length and we aimed to cover targets at 2×, resulting in a raw bait set of 40,269 probes. Probes were subsequently verified with the *Venustachoncha* genome and 10 of our TransRate-filtered transcriptomes (i.e. those of individuals E4, E8, E32, E36, Mol1-Mol6). The procedure consisted of repeated analyses with BLAST version 2.5.0+ (Camacho et al., 2009) to examine in silico how many hits these probes would produce on each genomic and transcriptomic reference under a variety of hybridization melting temperatures. After moderate filtering the probes were synthesized on magnetic beads for subsequent target enrichment.

2.7 | Preparation of genomic libraries

Target enrichment was performed with genomic DNA libraries of 96 specimens (95 Coelaturini and one iridinid bivalve), that is, 48 for continent-scale phylogenetics and 48 from six sampling localities (5–11 individuals per population) in the Malawi Basin for population genetics (Table S3). Two populations occur in the northern region, two in the southern region, one at Likoma Island and the last along the Shire River that drains Lake Malawi in the south. Genomic DNA was extracted from ~20 mg of dried tissue of the posterior or anterior adductor muscles, mantle tissue, and in case required the foot. DNA was extracted with the NucleoSpin 96 Tissue kit from Macherey-Nagel on a KingFisher Flex robot following manufacturer's recommendations. DNA concentrations in the extracts were quantified with Qubit fluorometry. We sheared DNA through sonication in a Bioruptor Pico: Samples with >20 ng/μl were subjected to 20 cycles of 30 s sonication, 30 s pause, for those with lower DNA concentrations we used between 18 and 3 cycles depending on DNA content. The results of library fragmentation were verified with chip-based capillary electrophoresis on an Agilent BioAnalyser. Genomic libraries were prepared with the NextFlex Rapid DNA-Seq version 2.0 kit of PerkinElmer and associated Unique Dual Indices for multiplexing.

2.8 | Target enrichment and sequencing

We validated our selection of targets, bait design and our molecular protocols with four target enrichment reactions. Pools were created at equimolar concentrations and enriched with a ~19 h incubation period at 65°C and 14–16 cycles in the post-hybridization PCR reaction. After enrichment, we purified the resulting products using 0.8 magnetic beads and rehydrated the enriched pool in 30 μl TLE. Post-enrichment libraries were quantified with Qubit and BioAnalyser, after which the results of multiple reactions were pooled at equimolar concentration and sequenced paired-end (2×150 bp) on an Illumina NextSeq 500 of the LIGAN-MP Genomics Platform of UMR8199.

2.9 | Bioinformatic analysis of sequencing results

2.9.1 | Open reading frames with HybPiper

Demultiplexed raw reads were cleaned with TrimGalore!, CutAdapt version 2.6 (Martin, 2011) and Trimmomatic version 0.39 (Bolger et al., 2014) performing iterative FastQC quality controls at each step. The cleaned paired-end reads of each specimen were subjected to the HybPiper workflow version 1.3.1 (Johnson et al., 2016). This workflow starts by mapping reads to our ORF targets and sorting them per ORF in separate directories with BLASTX and BWA version 0.7.17 (Li & Durbin, 2009) after removal of paired duplicated reads. Subsequently, the reads per ORF are subjected to de novo assembly using SPAdes version 3.14.0 (Bankevich et al., 2012) resulting in the construction of a supercontig per ORF and per specimen. The resulting contigs are sorted and the boundaries between exons and flanking regions are predicted with Exonerate version 2.4.0 (Slater & Birney, 2005). Alignments of the supercontigs, which include exons and intronic/intergenic flanking regions, were performed with MAFFT version 7.475 (Katoh & Standley, 2013) and MUSCLE, using our ORF targets to verify exon predictions.

2.9.2 | Ultraconserved elements with Phyluce

Demultiplexed and cleaned reads were also mapped to our UCE targets and treated with Phyluce. We assembled contigs with Trinity, which subsequently underwent orthology detection, paralogue removal and the matching of contigs to targets with LastZ (Harris, 2007), after which contigs were aligned with MAFFT or MUSCLE. Two essential parameters in the Phyluce pipeline are --min_identity and --min_coverage (--min_kmer_coverage was set to two), and we examined combinations of each factor between 50% and 80% at intervals of 5% (49 combinations in total). Many of these combinations are more permissive than the recommendations of Bossert and Danforth (2017), but we used additional downstream processing to eliminate potential contaminant sequences. Parameter combinations were evaluated using the proportion of unique contigs retrieved and the total number of UCEs retrieved for all 96 samples. The resulting data were analysed in R version 3.6.1 (R Core Team, 2019).

2.9.3 | Enrichment balance between ORFs and UCEs

To verify the enrichment balance between ORFs and UCEs, we examined the proportion of reads that mapped on either type of target. This task has been performed starting from the reads that were identified with HybPiper and Phyluce as mapping to ORF or UCE targets, respectively, which also provides information on the specificity of our enrichment strategy. More detailed information is obtained by extracting kmers of 18 to 21 bases from various

.fasta files with JELLYFISH version 2.2.10 (Marcais & Kingsford, 2011), that is, first from .fasta files that contain the ORF and UCE target sequences, and then for a sample of specimens in our data set. Each kmer that occurs in both the specimen and ORF dictionary is considered an ORF hit, whereas each kmer that occurs in the specimen and UCE dictionary is considered an UCE hit. The ratio of hits of both types indicates the enrichment balance between both sets of targets.

2.9.4 | Mitogenome skimming

Unionoida and some related bivalves have double uniparentally inherited mitochondria, which appears to be unique within the animal kingdom and which may play a role in sex determination (Breton et al., 2011; Zouros et al., 1994). Against this background, the abundant use of mitochondrial fragments in previous phylogenetic inferences for Unionidae (see above), and the potential role of conflict between mitochondrial and nuclear genes in speciation, we performed mitogenome skimming with a custom-build pipeline that outperformed MITOFINDER version 1.4.1 (Allio et al., 2020) for this specific data set. The tissues we used for genomic library preparation predominantly or exclusively contain maternally-inherited mitochondria (Breton et al., 2011; Froufe et al., 2020), so that recovering the paternally-inherited mitogenome falls beyond our current scope. First, we aligned reads on the mitochondrial genome of *Pygandon grandis* (NCBI GENBANK: NC_013661; Breton et al., 2009) with BOWTIE2 version 2.3.5 (Langmead & Salzberg, 2012) to produce a .bam file, which was sorted and indexed with functions of SAMTOOLS. Subsequently, the files for all individuals were merged with SAMTOOLS and visualized with IGV. This merged .bam file was converted into .fastq files after which the reads were assembled with SPADES. The resulting assembly was aligned to the *Pyganodon grandis* mitogenome and visually inspected to retain unambiguous contigs. The retained contigs were used to recover reads for 48 Coelaturini specimens from the Malawi Basin using again BOWTIE2, SAMTOOLS and SPADES, after which individual assemblies were reassembled with TRINITY. The total assembly was evaluated and the procedure was iterated each time with laterally extended starting contigs. Annotations were performed with MITOS (Bernt et al., 2013).

2.10 | Macroevolutionary analyses

2.10.1 | Alignment of UCE and ORF data

A single transcriptome was available in NCBI GENBANK for the subfamily Parreysiinae at the start of these analyses, namely that of *Scabies phaselus* (Lea, 1856) (SRX5281799; Pfeiffer et al., 2019), which was hence used as outgroup for phylogenetics. We used HYBPIPER as described above to recover information on our ORF targets for this outgroup. Phylogenetic analysis of the ORFs was performed using the supercontigs reconstructed for ingroup taxa with HYBPIPER. These supercontigs contain exons and intronic/intergenic flanking regions; alignments were produced with MAFFT. The UCEs were processed for phylogenetic analysis with PHYLUCE. Alignments for each UCE locus were obtained using phyluce_align_seqcap_align with MAFFT as aligner and without edge-trimming. Phylogenetic data sets of both UCEs and ORFs were filtered to retain target loci with >50% taxon completeness. ORF and UCE alignments were trimmed separately using BMGE version 1.2 (Cruscuolo & Gribaldo, 2010) with a maximum gap rate per sequence and character of 0.3 and a maximum entropy threshold of 0.4 to remove ambiguous regions. After concatenating trimmed single-locus alignments with AMAS version 1.0 (Borowiec, 2016), we used SPRUCEUP version 2020.2.19 (Borowiec, 2019) to detect outliers, which were replaced as missing data (windows size = 20, overlap = 15, criterion = lognorm, cutoff = 0.99). Alignment statistics were calculated with AMAS.

2.10.2 | Phylogenetic inference

We performed phylogenetic analysis with maximum likelihood (ML) on concatenated data sets for ORFs and UCEs separately to evaluate the congruence between both. We used AMAS to generate a partition file for the UCEs with the sliding-window site characteristics method based on site entropies (SWSC-EN; Tagliacollo & Lanfear, 2018) to define the limits of the UCE "core" and flanking regions. A single partition was considered for the supercontig of each ORF. Concatenated data sets and the partitioning information were subjected to phylogenetic inference in IQTREE version 2.0.3 (Minh et al., 2020), using the integrated MODELFINDER (Kalyaanamoorthy et al., 2017) to determine the best-fit substitution model for each partition. We used 1000 bootstrap replicates; branch support values were calculated with a Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT; Guindon et al., 2010) and ultrafast Bootstrap (UFBoot2; Hoang et al., 2018).

2.11 | Microevolutionary analyses

2.11.1 | Variant calling

Assessments of population genetic diversity and population structure are based on ORF targets without intronic/intergenic flanking regions, and UCEs with flanking regions. Variant calling was performed by alignment of the raw reads against custom-built consensus sequences for ORFs and UCEs with their respective flanking regions using BOWTIE2 on a sample-by-sample basis, after which .sam files were converted to .bam using SAMTOOLS, modified with PICARD version 2.21.4 (available at http://broadinstitute.github.io/picard), and merged across individuals with SAMTOOLS. These .bam files, one for ORFs, another for UCEs, were used to call and annotate single nucleotide polymorphisms (SNPs) with GATK version 4.1.9.0

(McKenna et al., 2010) using HaplotypeCaller. Individual GVCF files were merged per marker type and subjected to joint genotyping to obtain a .vcf file with information on all sites, both variant and invariant. From the .vcf file for ORFs we extracted information on exonic sites using an Exonerate-generated .gff annotation file. The resulting .vcf files were filtered with functions of VCFtools version 0.1.16 (Danecek et al., 2011) and BCFtools version 1.12 (Danecek et al., 2021) depending on downstream analyses as described below.

2.11.2 | Nucleotide diversity and population differentiation

For subsequent analyses of nucleotide diversity and population differentiation, the .vcf files for ORFs and UCEs were individually filtered to retain sites with a mean genotype depth of 8 that have been successfully genotyped in 80% of the individuals. Because invariant sites do not have quality scores when using condensed nonvariant blocks, we created individual .vcf files for variant and invariant sites. Invariant sites were identified by setting the minor allele frequency to zero (--max-maf), whereas variant sites have a minor allele count ≥1 (--mac). Only variant sites with a minimum quality score of 30 (-minQ) were retained. Subsequently, we indexed both .vcf files with tabix of SAMtools and combined them with BCFtools to estimate nucleotide diversity (π) within populations per ORF and UCE, the average, absolute nucleotide divergence between population pairs (D_{XY}) and population differentiation (F_{ST}) using Pixy version 1.0.4 (Korunes & Samuk, 2021). We also converted the filtered .vcf file for ORFs into a multifasta file with random (unphased) assignment of variants to the two alleles. Each multifasta file was aligned to its ORF target, after which reading frames were manually verified in SeaView. Gap-only sites in protein sequences were removed. The verified multifasta files were combined to calculate synonymous and non-synonymous nucleotide diversity (π_S and π_N, respectively) with dNdSpiNpiS version 1.0 (available at http://kimura.univ-montp2.fr/PopPhyl).

2.11.3 | Genetic structure

Analyses of genetic structure were undertaken for variant sites of ORFs only, which were obtained from the total .vcf file by filtering for a minimum allele count of 3, a minor allele frequency ≥0.05, a minimum quality score of 30 and a mean genotype depth of eight for at least 80% of the individuals. Furthermore, indels and variants that were not biallelic were removed, as were sites with a Hardy–Weinberg $p < .001$. Finally, we removed individuals for which >50% of the sites displayed missing data. The resulting .vcf file was converted into a .bed file for principal component analysis (PCA) using PLINK version 1.90b6.18 (Purcell et al., 2007), and to examine population structure with fastSTRUCTURE (Raj et al., 2014) using $K = 1–6$ with 30 independent runs per K. The underlying genetic structure and the appropriate number of clusters were examined with the ΔK method (Evanno et al., 2005) and others that are more robust when sampling is uneven (Puechmaille, 2016), as implemented in StructureSelector (Li & Liu, 2018).

3 | RESULTS

3.1 | RNA-seq and transcriptome assembly

Transcriptome sequencing resulted in on average 42,928,473 ± 14,030,906 paired-end reads with a GC content of 38.2 ± 1.6%. De novo assembly statistics are illustrated in Table 1. Filtering with TransRate significantly increased the quality of assemblies, as evidenced by substantially less duplicated BUSCO hits (Table 2; two-sample Wilcoxon Rank Sum test: $W = 199$, $p < .001$), but it decreased completeness. Raw Agalma assemblies have significantly more complete BUSCO hits ($W = 198$, $p < .001$), less fragmentary hits ($W = 31$, $p < .001$) and less missing data ($W = 8.5$, $p < .001$). However, we observed no significant difference between the raw Agalma assemblies and TransRate-filtered assemblies in the number of complete single-copy BUSCO hits ($W = 132$, $p = .436$). The clustered supertranscriptome contained 988,460 contigs in total and BUSCO analysis against the Metazoa_odb9 database indicated that it is very complete, but with a high level of redundancy. Prediction and clustering of ORFs resulted in the retention of 131,503 ORFs, and effectively decreased the number of duplicated BUSCO hits and therewith redundancy (Table 2; Ortiz-Sepulveda et al., 2022).

3.2 | Selection of ORF target regions

Mapping the clustered ORFs from our supertranscriptome to the BUSCO Metazoa_odb9 database, we retained 633 single-copy and 334 duplicated BUSCO hits, respectively (Table 2, last column), of which 633 and 186, respectively, were retained as likely single-copy orthologous targets across our ingroup taxa. Evaluation of orthology for ORFs that mapped to the Unioverse probe set suggested that 186 of the 811 Unioverse loci (22.9%) are affected by homology issues for our ingroup taxa (which belong to the taxa for which the Unioverse probe set was designed). In most of these cases, several divergent ORFs mapped to a single Unioverse locus, suggesting paralogy, but we also observed instances where Unioverse loci were not orthologous to their "associated" *Bathymodiolus* target region, as indicated by less sequence divergence between *Bathymodiolus* and the matching fragment of our ingroup ORFs than between *Bathymodiolus* and the associated Unioverse loci. Nevertheless, our evaluation suggested most Unioverse loci to be single-copy orthologous in Coelaturini, which resulted in the addition of 297 ORF targets from the Unioverse probe set (usually several Unioverse loci map to a single ORF). Mapping ORFs and subregions among each other resulted in the removal of one ORF from the duplicate BUSCO selection and another from the Unioverse set, resulting in a total of 1114 retained ORFs which cover 1,677,936 nucleotides (on average 1506 nt/ORF).

TABLE 1 Statistics on de novo transcriptome assembly for 12 selected Coelaturini specimens: E4, E32, *Coelatura* (*Nyassunio*) *nyassaensis*; E8, E27, E30, E36, *Coelatura hypsiprymna*; Mol1, Mol2, *Grandidieria burtoni*; Mol3, Mol6, *Nitia acuminata*; Mol4, Mol5, *Coelatura hauttecoeuri*

RNA_ID	#contigs	Smallest	Largest	#bases	mean_len	n > 1 k	n > 10 k	n_with_orf	mean_orf_%	n90	n70	n50	n30	n10	GC
E4	214,034	301	14,438	223,902,463	1046.11	65,224	106	42,978	37.27	417	873	1678	2774	4835	0.350
E8	213,681	301	28,564	274,210,922	1283.27	78,427	513	49,493	36.37	476	1222	2288	3654	6333	0.348
E27	165,991	301	26,557	203,347,366	1225.05	58,336	425	45,402	43.40	454	1140	2138	3481	6417	0.360
E30	168,380	301	21,779	198,609,002	1179.53	60,021	94	43,638	39.28	452	1087	1989	3127	5225	0.352
E32	143,592	301	26,221	173,857,884	1210.78	51,543	270	40,092	42.57	457	1127	2052	3292	6032	0.357
E36	155,099	301	27,892	199,983,363	1289.39	58,649	322	43,364	40.77	480	1247	2259	3575	6166	0.356
Mol01	137,268	301	20,075	154,440,102	1125.10	46,734	125	34,257	42.01	443	994	1821	2892	5028	0.358
Mol02	115,197	301	26,274	126,001,086	1093.79	37,633	153	31,015	45.52	437	938	1738	2810	5064	0.363
Mol03	192,418	301	25,075	253,191,365	1315.84	72,003	649	47,610	39.13	484	1276	2375	3768	6649	0.354
Mol04	115,433	301	13,112	114,159,368	988.97	35,909	28	29,696	45.32	419	824	1462	2288	3814	0.362
Mol05	235,569	301	33,921	303,293,623	1287.49	83,497	916	53,363	39.01	467	1225	2359	3861	7081	0.354
Mol06	184,405	301	26,842	261,822,189	1419.82	72,120	1067	48,094	41.34	512	1433	2620	4216	7766	0.356

Note: N > 1 k and n > 10 k indicate the number of contigs larger than 1000 nt and 10,000 nt, respectively; mean_orf_% indicates the proportion of contigs contained within open reading frames. Values for n90, n70, n50, n30, n10 indicate the length of the 90%, 70%, 50%, 30%, 10% longest contigs in the assembly.

TABLE 2 Comparison of BUSCO benchmark analyses to evaluate the completeness of our raw AGALMA and TRANSRATE-filtered assemblies for our 12 Coelaturini transcriptomes, the resulting clustered supertranscriptome (SupTr) with its raw predicted ORFs (ORF) and the resulting homologous ORF clusters (Cl_ORF). Results for analyses of raw AGALMA assemblies are represented at the top, whereas those filtered with TRANSRATE for assembly errors at the bottom. Taxon identifications are as in Table 1

	E4	E8	E27	E30	E32	E36	Mol1	Mol2	Mol3	Mol4	Mol5	Mol6	SupTr	ORF	Cl_ORF
Raw AGALMA assemblies															
Complete BUSCO	919	963	947	940	938	951	931	925	967	892	963	965	976	965	967
Complete-singleCopy	560	510	544	525	572	547	563	588	543	589	483	526	71	68	633
Complete-duplicated	359	453	403	415	366	404	368	337	424	303	480	439	905	897	334
Fragmented	49	12	20	28	26	19	33	40	5	55	7	5	1	4	1
Missing	10	3	11	10	14	8	14	13	6	31	8	8	1	9	10
TRANSRATE-filtered assemblies															
Complete BUSCO	796	470	633	457	438	486	623	745	405	618	517	504	965	953	953
Complete-singleCopy	648	431	549	423	421	457	552	644	387	580	449	468	257	268	755
Complete-duplicated	148	39	84	34	17	29	71	101	18	38	68	36	708	685	198
Fragmented	84	90	89	106	85	83	97	79	67	104	87	68	2	11	11
Missing	98	418	256	415	455	409	258	154	506	256	374	406	11	14	14

3.3 | Selection of ultraconserved elements

Iterative inclusion of additional genomes resulted in the selection of 553, 1649 and 4040 UCEs that are shared among seven, six and five genomes, respectively (any set of 5 genomes includes both bivalve and gastropod genomes). After filtering candidate UCEs to mask regions with repeated motives, low-level complexity or a length <100 nt, we retained 388, 1257 and 3366 UCEs (5011 in total) for seven, six and five genomes, respectively. Further filtering and overlap detection (within UCEs and to ORFs) resulted in the retention of 4107 UCEs, which jointly cover 595,060 nt.

3.4 | Evaluation of target regions

Of the 84,484 ART_ILLUMINA reads, 66,804 mapped onto the *Venustaconcha* genome (79.1%), all of which were unique hits. These hits covered 3931 of our 5221 target regions (75.3%), that is, 631 of 633 (99.7%) of our complete single-copy BUSCO ORFs, all of our complete duplicate BUSCO ORFs, 295 of 296 (99.7%) UNIOVERSE ORFs, 996 of the 1255 stringent UCEs (79.4%; i.e., UCEs found in at least 6 genomes) and 1824 of our 2852 less stringent UCEs (64.0%; UCEs found in 5 of our 7 genomes), indicating a higher efficiency for ORFs than UCEs. This mapping indicated that ORFs were regularly (but not always) retrieved completely on the same *Venustaconcha* contig, and that we could expect to retrieve multiple exons per ORF. In the *Venustaconcha* genome these exons were typically larger than 200 nt and often separated from other exons of the same ORF by 1000 or 10,000 s of nt.

Our targets for bait design covered 5221 genomic regions with a length of 2,272,996 nt. Of the 40,269 raw probes, 37,959 passed quality control (~94.3%). The impact of this filtering on the overall coverage of our target regions was minimal; however, for three UCEs we were not able to develop any probes and for four ORFs the discarded probes resulted in gaps of >300 nt, so that these ORFs were expected to be incompletely covered upon target enrichment.

3.5 | Recovery of open reading frames and ultraconserved elements

On average, we obtained over 3 million reads per sample (range: 35,246–9,132,732), of which (mean ± SD) 61.81 ± 13.94% were on target. Of these on-target reads, 60.83 ± 14.15% relate to ORFs whereas 0.97 ± 0.73% to UCEs. There was a weak but significant positive correlation among the total number of reads per sample and the proportion of on-target reads ($R^2 = .045$, $F = 4.442$, df = 1 + 94, $p = .038$), but no trend in the total number of reads per phylogenetic clade within Coelaturini ($R^2 = .020$, $F = 1.961$, df = 1 + 94, $p = .165$; Figure S1). We observed unbalanced enrichment of UCEs versus ORFs: Whereas the UCE regions contain 26.17% of the total of targeted nucleotides, the number of UCE kmer hits compared to ORF kmer hits is around 0.05%, indicating a substantial underrepresentation of UCEs compared to ORFs.

On average over 1102 of the 1114 ORFs were consistently enriched and mapped for all 95 unionids (857 are consistently recovered for over 50% of their length in all specimens), with the exception of the distant iridinid specimen (dna0240; see Figure 2). HYBPIPER detected hidden paralogy in at most 2 ORFs per specimen. As the number of reads obtained for a sample decreases, we see a gradual decrease in the recovery of ORFs, which becomes more marked for samples with <500,000 reads ($n = 12$). As to UCEs, we recovered data for up to 1905 out of the 4104 UCEs (46.5%), and the coverage per sample was proportional with the number of reads (linear model: $r^2 = .557$, $p < .001$), as was observed for the ORFs (Figure 2). The combination of 55% and 60% thresholds on sequence coverage and identity, respectively, maximized the total yield over all specimens, but it decreased the number of retrieved UCEs slightly to 1895. On average 281 UCEs are covered per individual (range 30–473; total of 26,982 regions recovered for 96 samples). The number of recovered UCEs, and the proportion of unique contigs recovered per individual decreased gradually as the thresholds on sequence coverage and identity were altered, with more abrupt decreases when the threshold on % identity was increased to ≥70% (Figure 3). The consistency with which UCEs are recovered across taxa is low: 37 and 276 UCEs are recovered in >75% and >50% of individuals, respectively. The length distribution of the retained UCEs is highly similar to that of all UCEs (Figure S2).

3.6 | Mitochondrial genome skimming

Iterative rounds of genome skimming on 48 specimens allowed us to reconstruct the entire maternally-inherited mitogenome for Coelaturini from the Malawi Basin. This mitogenome contains 15,664 bp, and annotation included the 37 expected genes (Boore, 1999): 2 for rRNA, 22 for tRNAs, which encode components in the mitochondrial translational machinery, and 13 other genes that encode protein components of the respiratory chain and ATP synthase (Figure 4). On average ~40% of the mitogenome was recovered per individual (range 0%–100%), with consistent recovery of several gene regions across most individuals, which would enable integration of mitogenomic data in phylogenetic and population genetic analyses.

3.7 | Macroevolutionary analyses

All 95 Coelaturini specimens were included in our phylogenetic analyses. Selection of loci with >50% taxon-completeness resulted in a data set that contained 1109 ORF supercontigs with a mean and total alignment length of 2118 and 2,348,614 bp, respectively. For UCEs, the cleaned and trimmed data set contained 276 loci with a mean and total alignment length of 432 and 119,105 bp, respectively. For these,

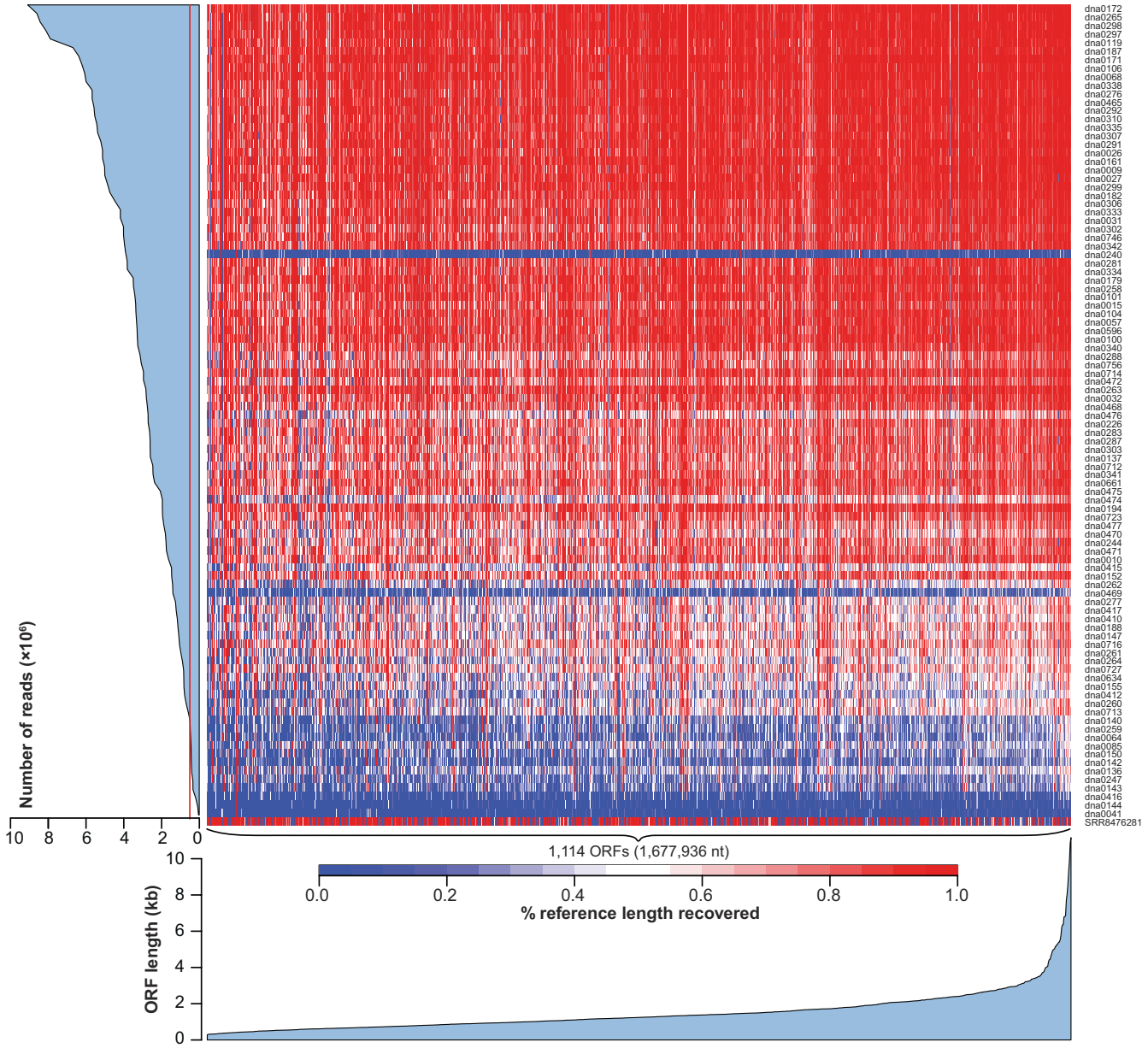

FIGURE 2 Percentual recovery of the total length of 1114 ORFs for 96 sequenced taxa and an outgroup transcriptome. The total size of ORFs, on average ~1500 nt, and the number of million filtered reads per sample are also indicated. The red line as to number of reads indicates a cutoff value of 500,000 reads. Samples with a lower number of reads displayed a drastic decrease in ORF recovery. A drastic decrease in recovery was also observed for the outgroup sample dna0240 belonging to the family Iridinidae, despite having generated 3,900,000 clean reads.

515,219 (21.9%) and 11,001 (9.24%) sites, respectively, were phylogenetically informative. Additional data on specimen representation, alignment length, and the proportion of informative and missing sites for each ORF and UCE locus are given in Ortiz-Sepulveda et al. (2022). The partition analysis resulted in a 402-partition scheme with 45 unique substitution models for ORFs, and a 155-partition scheme and 51 unique substitution models for UCEs (Ortiz-Sepulveda et al., 2022). The phylogenetic trees reconstructed from the ORF and UCE data sets separately are highly congruent (Figure 5; Figures S3 and S4). Whereas the ORF tree is fully supported, except for at population-level branches within the "Malawi" clade, some uncertainty exists as to terminal branches in the UCE tree, as evidenced by decreasing support and some branch rearrangements at very shallow levels of divergence (Figure 5; Figures S3 and S4).

3.8 | Microevolutionary analyses

For each ORF and UCE we genotyped 48 diploid, dioecious individuals belonging to six populations from the Malawi Basin. Filtering for nucleotide diversity and population differentiation resulted in the retention of 1097 ORFs and high-quality genotypes for

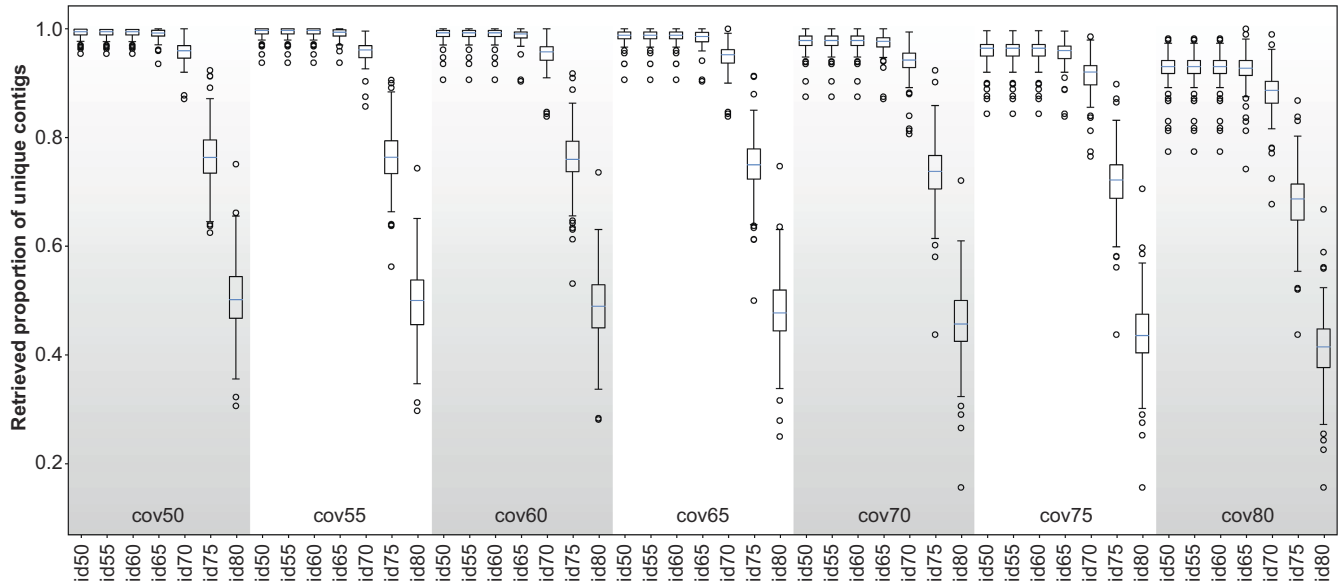

FIGURE 3 UCE recovery depending on combinations of coverage and identity as specified in the PHYLUCE pipeline. Boxplots indicate the contig ratio for all individuals, that is, the number of unique contigs/maximum number of unique contigs per individual for all 96 individuals. The total number of UCEs recovered for 96 individuals varied between 1905 (combination cov50 and id50-60) and 926 (combination cov80 and id80). Several scenarios where coverage and identity are between 50% and 65% resulted in similar results, whereas both the number of unique contigs and UCE recovery decrease substantially for scenarios with identity >70%.

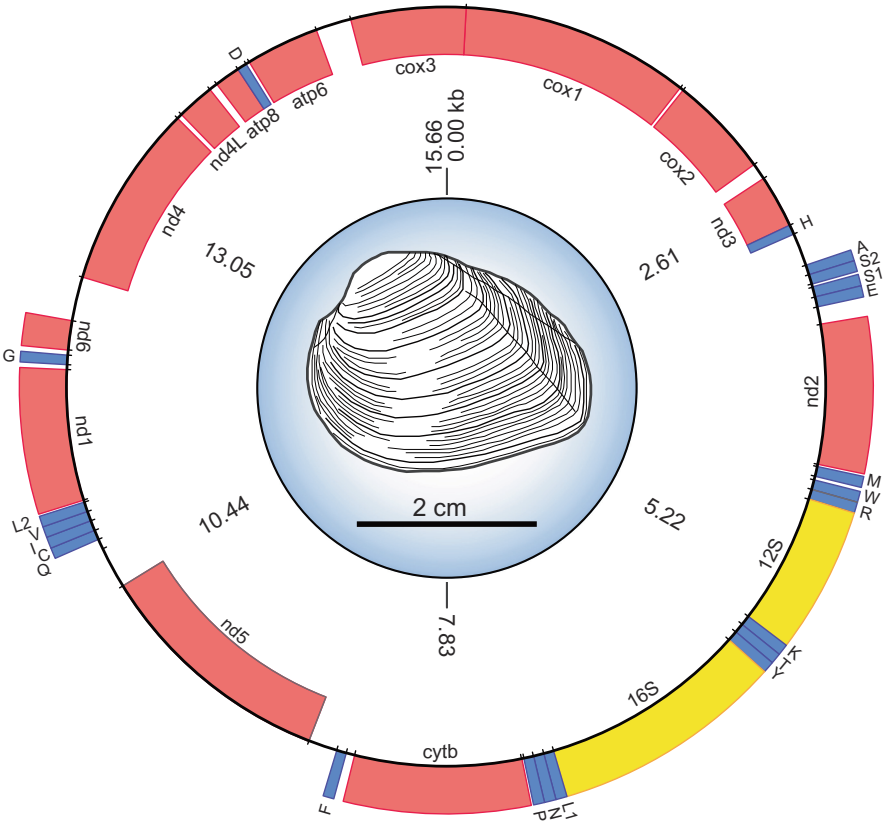

FIGURE 4 Gene map of the maternally-inherited mitochondrial genome for Coelaturini from the Malawi Basin. Genes positioned on the 5′ to 3′ (positive/heavy) strand are indicated on the inner circle, whereas those on the 3′ to 5′ (negative/light) strand on the outer circle.

~1,153,827 out of 1,675,638 sites (68.9%), of which 17,988 were polymorphic (1.1%; Ortiz-Sepulveda et al., 2022). Using identical filtering criteria, we retained 309 UCEs with high-quality genotypes for 111,492 out of 254,694 sites (43.8%), of which 3148 were polymorphic (1.2%). Estimates of π average 0.00167 ± 0.00316 and 0.00116 ± 0.00206 for ORFs and UCEs, respectively, and they are very similar across the six studied populations (Figure 6a; Figure S5a), indicating an average of ~1915 and ~130 nucleotide differences in pairwise haplotype comparisons, respectively. The average π_S per population varies between 0.00220 and 0.00329,

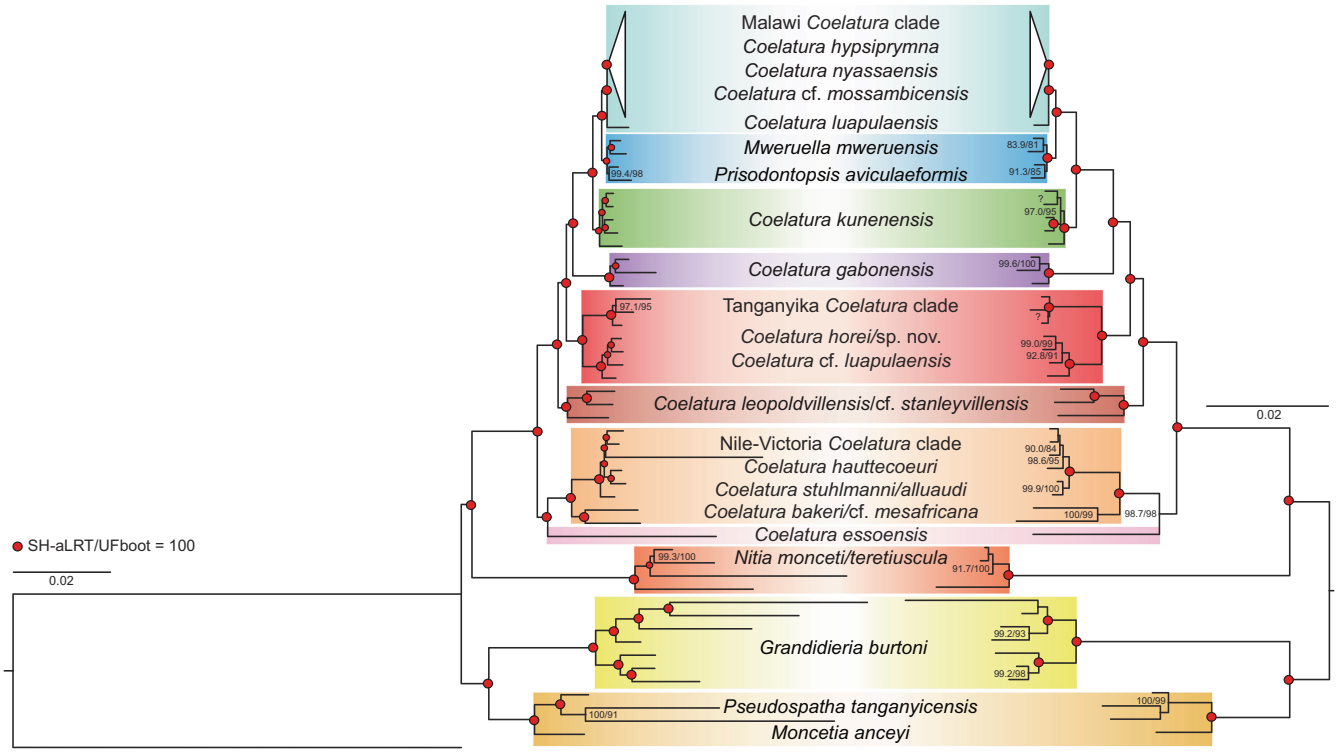

FIGURE 5 Maximum likelihood phylogeny of Coelaturini based on a concatenated data set of 1109 open reading frames (2,348,614 bp with 515,219 parsimony informative sites; left), including exons and their intronic/intergenic flanking regions, and based on a concatenated data set of 276 ultraconserved elements (119,105 bp with 11,001 parsimony informative sites; right). Red nodes are fully resolved using a Shimodaira-Hasegawa approximate likelihood ratio test and ultrafast bootstrapping. Both data sets result in highly congruent topologies, which are fully or mainly supported for ORFs and UCEs, respectively. More details are provided in Figures S3 and S4.

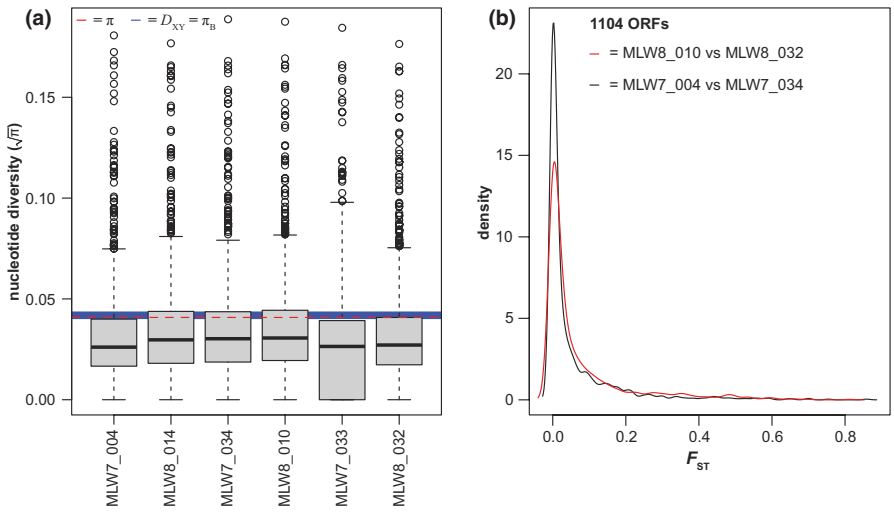

FIGURE 6 (a) Square root transformed nucleotide diversity (π) within six populations of Coelaturini from the Malawi Basin as inferred from ORFs (without intronic/intergenic flanking regions). The nucleotide diversity averaged over all six populations is indicated with a dashed red line, whereas mean population pairwise sequence divergence for each of 15 population pairs, that is, mean, square root transformed D_{XY}-values, are indicated with blue lines (they are highly similar for all population pairs, resulting in a bold blue line). (b) the density distribution of population pairwise F_{ST}-values for 1104 ORFs (one value per ORF) for two out of the 15 population pairs (these distributions are also representative for other population pairs).

whereas π_N varies between 0.00057 and 0.00092, resulting in an average π_N/π_S of 0.259 (Figure S6). Including intronic/intergenic flanking regions of ORFs would further increase the number of variant sites that can be analysed. Pairwise D_{XY}-values average 0.00175 ± 0.00008 for ORFs and 0.00127 ± 0.00015 for UCEs, which is only 5%–9% higher than the mean π, respectively

(Figure 6a; Figure S5a), indicating limited overall net nucleotide divergence among populations. F_{ST}-values average 0.060 ± 0.019 for ORFs and 0.054 ± 0.015 for UCEs, indicating moderate genetic differentiation. Substantial variation exists in F_{ST}-values among ORFs (Figure 6b) and UCEs (Figure S5b). Per pairwise population comparison between 60 and 230 ORFs and between 16 and 38 UCEs display F_{ST}-values >0.15, of which 42.0% and 59.3%, respectively, display elevated D_{XY}-values too (i.e., $D_{XY} > 0.002$).

Filtering ORF data to examine genetic structure resulted in the removal of two individuals (dna0469 and dna0416) and a final data set of 2161 SNPs (Ortiz-Sepulveda et al., 2022). PCA on this data set indicated that PC1, 2 and 3 represent 11.7%, 10.9% and 8.7% of all variation in the data set, respectively. The 95% convex hulls of populations overlap substantially within the northern and southern regions, but not between them (Figure 7). Both regions are mainly separated along PC3. The Likoma Island population falls closer to populations of the northern region in PC1 versus 2, but closer to those of the south in PC1 versus 3. The population of the Shire River overlaps with one population from the south, but shows substantial differentiation from the other southern population. These results are highly congruent with those obtained with fastSTRUCTURE on the same data set, which suggest $K = 4$ to be the best scenario with the ΔK method and most of the estimators of Puechmaille (2016). Some of these latter estimators suggested five clusters, but with specimen assignments that are almost identical to the $K = 4$ solution (Figure 7). Two of these four clusters correspond to sampling locations, that is, Likoma Island and Shire River, whereas the others coincide with a north–south separation in which one population from the south (MLW8-014) displays mixed assignments, including signatures from the northern and Shire River clusters. Interestingly, the Shire River cluster, although being geographically in the far south, clusters with the north in the $K = 3$ scenario.

4 | DISCUSSION

4.1 | Target selection and capture

Our screening illustrates the robustness of identifying ORFs that are probably single-copy and orthologous from existing databases followed by paralogue detection based on ingroup transcriptomes. Our enrichment strategy performed well and was not affected much by the unknown intron-exon boundaries upon probe design. This result opens opportunities to use ORF predictions from transcriptome assemblies in other Metazoa as direct targets for probe design, provided orthology is verified. Therewith, our strategy simplifies the development of genomic data sets significantly, especially for non-model organisms. If many short exons are expected, the use of shorter probes, for example, of 80 nt, or covering targets more densely with probes, for example, at 3× or 4×, may further enhance the capture efficiency.

The effectiveness of selecting UCEs for unionids with the PHYLUCE pipeline is somewhat hampered because few and only distant genomes were available for molluscs (Sigwart et al., 2021; Sun et al., 2019) compared to other taxa for which UCE sets have been developed. Nevertheless, we recovered hits for 1895 (46%) of our target UCEs, which is comparable to values obtained in some previous UCE studies (e.g., Kulkarni et al., 2020; Starrett et al., 2017; Streicher et al., 2018), indicating that our design worked. As is regularly the case in UCE studies (Buenaventura et al., 2021; Faircloth et al., 2012; Kulkarni et al., 2020; Quattrini et al., 2018; Starrett et al., 2017), the number of UCEs that can eventually be included in the alignment for phylogenetic inference was restricted to a subset of UCEs with high recovery across all ingroup taxa (but see Branstetter et al., 2017). Phylogenetic analysis on 276 UCEs allowed to unambiguously reconstruct the backbone phylogeny of Coelaturini and estimates of population genetic diversity from 309 UCEs were comparable to those obtained from ORFs, but more similar to the diversity at non-synonymous than at synonymous sites.

Combining ORFs and UCEs in the same probe set has resulted in competition: Although UCEs account for over 25% of the probes, only ~1% of our reads cover UCE targets. A potential factor of influence is the phylogenetic distance among the genomes used to identify UCEs and our ingroup, compared to the selection of ORFs based on ingroup transcriptomes. The recovery of 1895 UCEs across our samples despite having only ~1% of our reads mapping to UCE targets indicates that the issue results from hybridization efficiency rather than probe design, however. This result was unexpected based on a previous integration of multiple types of markers (Hutter et al., 2019), where no such competition was observed, but in that study the average length of UCE targets was >700 bp, compared to ~145 bp in ours. As we did not find a relationship between the length and recovery of UCEs (Figure S2), the most likely explanation is that differences in inherent properties of UCE and ORF targets (e.g., mismatches to genomic libraries or differences in melting temperatures) cause variation in sensitivity and specificity during hybridization. This hypothesis is corroborated by the more restricted recovery of UCEs compared to ORFs upon in silico mapping of reads onto the *Venustaconcha* genome. UCE recovery could be enhanced by altering the temperatures of hybridization and washing reactions, but further work is required to better understand the balance of enrichment across UCEs and ORFs.

4.2 | Mitochondrial genome skimming

As cells contain multiple copies of the mitogenome versus one copy of the nuclear genome, enrichment bias can be expected upon including both nuclear and mitochondrial targets in joint enrichment. Considering this issue, we aimed to recuperate mitochondrial data via genome skimming. We recovered on average ~40% of the

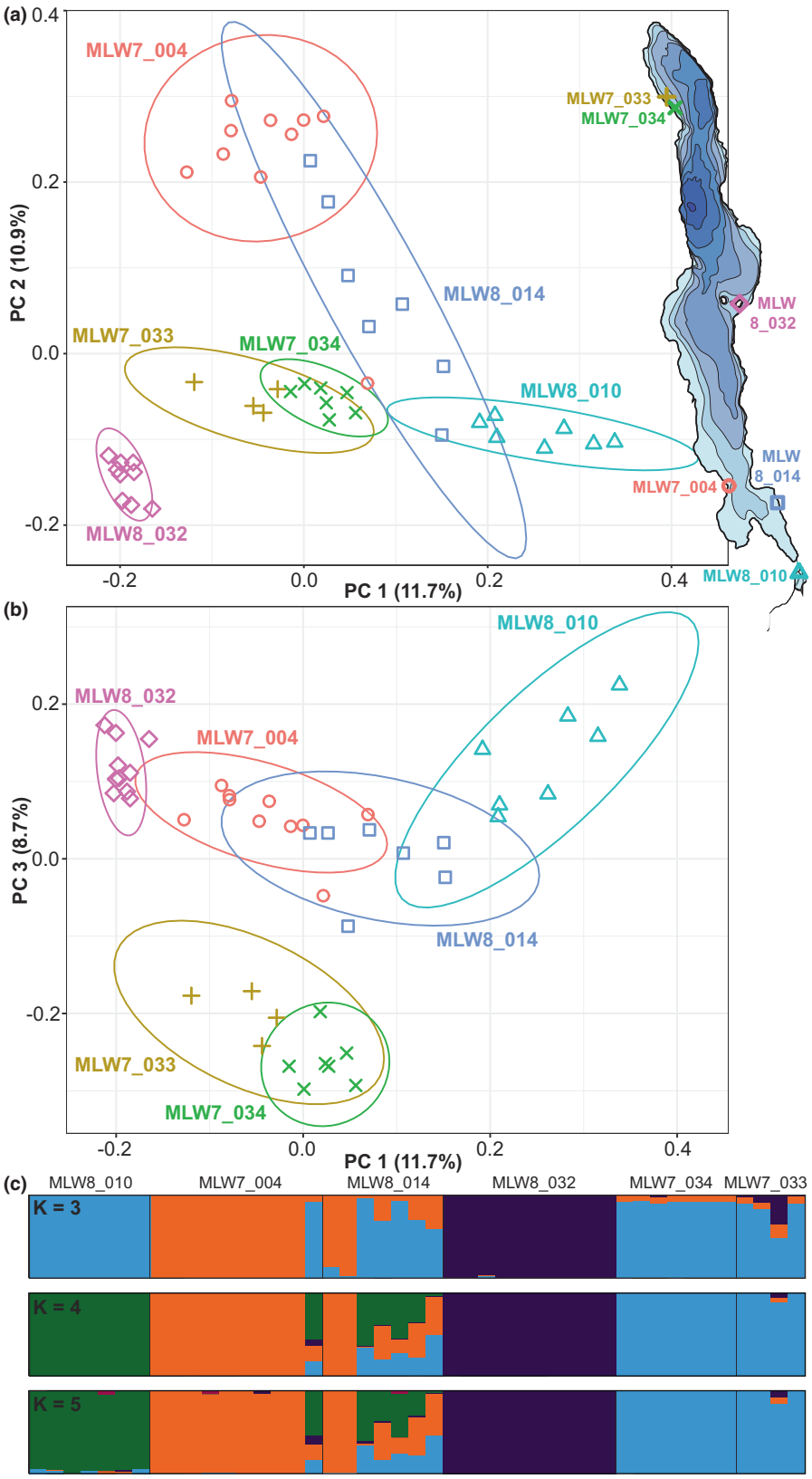

FIGURE 7 Structure of molecular diversity in Coelaturini from the Malawi Basin. (a, b) Principal component analysis on genome-wide SNP data, with 95% convex hulls on sampling localities. A bathymetric map of Lake Malawi, its outflow and the studied populations is provided in the inset. (c) Bayesian clustering with fastSTRUCTURE on the same SNP data set returned most support for a four-cluster solution separating the northern and southern regions of the Malawi Basin, and additionally the populations of Likoma Island (MLW8_032) and the Shire River (MLW8_010).

length of the mitogenome per specimen with a depth between 1× and 300×, which allowed us to reconstruct the entire maternally-inherited Coelaturini mitogenome (note that the tissues expressing the paternally-inherited mitogenome were not sampled here). This mitogenome revealed the type UF1 gene order, which would have characterized the most recent common ancestor of Unionidae and which is conserved in the subfamilies Ambleminae, Unioninae, but not in Gonideinae (Froufe et al., 2020). No members of Parreysiinae have been subjected to mitogenomic analyses, but our results show that the UF1 gene order prevails in this subfamily, or at least in

Coelaturini. Beyond the position of tRNA H and D, this gene order is also highly similar to that of the unionid male mitogenome (Lopes-Lima, Fonseca, et al., 2017).

4.3 | Macroevolutionary analyses

Despite important differences in the enrichment efficiency of ORFs and UCEs, data sets of each marker type produced highly similar phylogenies of Coelaturini, which allowed to resolve previous ambiguities (Ortiz-Sepulveda et al., 2020), both along deep and shallow phylogenetic branches. Examples include that *Grandidieria burtoni* is recovered as the sister clade to *Pseudospatha tanganyicensis* and *Moncetia anceyi* rather than to all Coelaturini, and that *Coelatura luapulaensis* is sister to the "Malawi" clade, instead of being part of this clade as previously recovered. Additional samples indicate that *Coelatura* from West Africa represents the sister-group to *Coelatura* from the Nile and Lake Victoria. At shallow nodes topologies diverged somewhat between the ORF and UCE trees, with an associated decrease in support values. Our phylogenies also indicate that *Grandidieria burtoni*, *Pseudospatha tanganyicensis* and *Coelatura* spp. from Lake Tanganyika represent species complexes, as was already presumed by Ortiz-Sepulveda et al. (2020).

4.4 | Microevolutionary analyses

Estimates of nucleotide diversity, pairwise population sequence divergence and F_{ST}-values were similar for ORF and UCE data sets, suggesting general robustness. Synonymous nucleotide diversity was overall low for the "Malawi" clade (average $\pi_S = 0.00271$) compared to species-level estimates from transcriptomic data across molluscs ($\pi_S = 0.02878$), or Metazoa altogether ($\pi_S = 0.01912$; see Romiguier et al., 2014). Our estimates of π_N/π_S were high compared to those of other Metazoa but similar to those of certain Darwin finches and Echinodermata (see Figure S6; Leroy et al., 2021; Romiguier et al., 2014). This low diversity may relate to the brooding ability and parasitic life-cycle of *Coelatura* species, but also to Late Pleistocene ecological crises in the Malawi Basin, which may have caused population bottlenecks in the aquatic fauna (Cohen et al., 2007; Ivory et al., 2016), and which may explain high π_N/π_S-ratios. Beyond low nucleotide diversity, also the pairwise nucleotide divergence among sampling localities is low, which is very similar to what has been observed in Malawi cichlids (Malinsky et al., 2018), as in recent speciation in *Ficedula* flycatchers (Ellegren et al., 2012). In Malawi cichlids the distributions of individual heterozygosity at nucleotide sites and of pairwise nucleotide divergence between species are partially overlapping, and multiple radiative events are interconnected by gene flow (Malinsky et al., 2018). Malawi Coelaturini display a similar pattern when comparing nucleotide diversity and pairwise nucleotide divergence (Figure 6a), which may be driven by similar population histories due to ecological interactions, as Coelaturini have a fish-parasitizing larval stage, or by common environmental change. F_{ST}-values indicate moderate genetic differentiation comparable to that in the *Lanistes* gastropod radiation of the Malawi Basin (Van Bocxlaer, 2017). Interestingly, several ORFs and UCEs for each pairwise population comparison displayed both elevated F_{ST} and D_{XY} values, suggesting that such F_{ST} values do not result from local reductions of genetic diversity in the genome (Charlesworth, 1998), but rather from diversifying selection. Despite the low degree of differentiation in Coelaturini, analyses of geographic structure differentiate several gene pools with a similar geographic pattern to that in *Lanistes* (Van Bocxlaer, 2017). Geographic differentiation between the northern and southern regions of the Malawi Basin is observed, however, the geographic structure in *Lanistes* is more clearly delineated than that in Coelaturini. Further population samples are required to examine the demographic history and population divergence of Coelaturini from the Malawi Basin. Nevertheless, our results seem compatible with an early stage of speciation (see Seehausen et al., 2014). If generalizable to a macroevolutionary scale, this feature may have contributed to the lack of resolution within geographic clades in previous phylogenetic studies (Ortiz-Sepulveda et al., 2020). We are hopeful that the enrichment and analytical strategy described here will enable cost-effective in-depth studies of genetic diversity and divergence, and therewith, an integration of diversification dynamics at micro- and macroevolutionary scales (insights into costs associated with molecular biology and sequencing are provided in Table S4).

4.5 | Advantages and disadvantages of our workflow

The here presented workflow provides good perspectives to enrich entire ORFs and intronic/intergenic flanking regions without prior knowledge of exon-intron boundaries, thus in the absence of a proximate well-assembled and annotated reference genome. Combining the selection of candidate ORFs using existing databases with orthology assessment from ingroup transcriptomes provides a useful approach for non-model organisms, that moreover allows leveraging museum specimens as long as some fresh samples across the ingroup are available. Our verification includes target alignment and manual verification, as recommended previously (Teasdale et al., 2016), which provides empirical scientists a trackable connection to their high-throughput sequencing data. As our approach features sequence capture, it is flexible, repeatable, and it allows the inclusion of targets from previously developed probe set, as demonstrated here with the Unioverse probe set (see Pfeiffer et al., 2019). Therefore, it enhances opportunities to effectively expand and reuse already published data sets. Our procedures indicate good recovery of ORFs, but if one aims to integrate UCE and ORF targets in the same enrichment reactions additional verifications to balance such reactions are recommended. Despite the reduced alignment length compared to ORFs or the smaller pool of SNPs, our retained UCEs contain substantial phylogenetic and population genetic information. At both scales, analyses based on ORFs and UCEs

produced highly comparable results, suggesting that our genomic sampling is representative. Estimates of nucleotide diversity for UCEs were closer to those at non-synonymous than at synonymous sites of ORFs, indicating that UCEs and their flanking regions are under selective constraints rather than being neutrally-evolving. Decisions on whether or not to include multiple marker types in the same enrichment strategy strongly depend on the questions to be addressed (see Hendriks et al., 2021). Whereas certain questions may adequately be answered using a single marker type, more representative sampling across the genome increases opportunities to reliably document evolutionary patterns, including the degree of phylogenetic congruence or the robustness of population-level summary statistics, and, therefore, it may enhance comparability across taxa. Furthermore, mitochondrial (or chloroplast) genomes may be recovered by skimming off-target reads, albeit with more variable sequencing depth.

AUTHOR CONTRIBUTIONS

Bert Van Bocxlaer and Xavier Vekemans designed research; Bert Van Bocxlaer, Claudia M. Ortiz-Sepulveda, Christian Albrecht collected biological materials; Claudia M. Ortiz-Sepulveda, Christelle Blassiau, Cécile Godé performed molecular biology; Mathieu Genete, Claudia M. Ortiz-Sepulveda, Bert Van Bocxlaer performed bioinformatic analyses; Bert Van Bocxlaer selected target ORFs and verified orthology; Mathieu Genete selected UCEs; Bert Van Bocxlaer performed genome skimming; Claudia M. Ortiz-Sepulveda performed phylogenetic and population genetic analyses with the help of Bert Van Bocxlaer and Xavier Vekemans; Bert Van Bocxlaer drafted the manuscript with help of Claudia M. Ortiz-Sepulveda and Xavier Vekemans; Bert Van Bocxlaer and Xavier Vekemans acquired funding.

ACKNOWLEDGEMENTS

We thank Heinz Buscher, Kevin Cummings, Daniel Engelhard, Thies Geertz, Daniel L. Graf, Adrian Indermauer, Elena Jovanovska, Alidor Kankonda, Charles Lange, Nicolas Lichilin, Ellinor Michel, Papy Mongindo, Idrissa Ouedraogo, Frank Riedel, Walter Salzburger, Friedemann Schrenk, Danny Simbeye, Ernest Tambwe, Jonathan A. Todd, Julius Tumusiime, Daan Vanhove, Emmanuel Vreven, Raphael Wangalwa and Oscar Wembo for providing samples or help with fieldwork. Brian Brunelle of Arbor Biosciences designed the probes and provided guidance on enrichment reactions. Sylvain Santoni and Julien Derop advised on RNA-seq and Illumina NextSeq procedures, respectively. We thank Vincent Castric, Sophie Gallina, Sylvain Legrand, and Camille Roux for discussion. Our study was substantially improved following constructive criticism by R. Nicolas Lou, John Pfeiffer, three anonymous reviewers and the subject editor, Andrew DeWoody. This study was funded by ANR-JCJC-EVOLINK (to BVB), FEDER-ERC-EVORAD (to BVB) and it contributes to the CPER research project CLIMIBIO (XV). We thank the French Ministère de l'Enseignement Supérieur et de la Recherche, the Agence Nationale de la Recherche, the European Fund for Regional Development (FEDER) and the region Hauts-de-France (HdF). We also acknowledge the UMR 8199 LIGAN-MP Genomics platform (Lille, France) of the "Federation de Recherche" 3508 Labex EGID (European Genomics Institute for Diabetes) funded by ANR-10-LABX-46, ANR-10-EQPX-07-01, FEDER, HdF.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The generated genomic data have been made available via the SRA of NCBI GenBank under BioProject PRJNA893605, including 108 BioSamples. BioSamples SAMN31437443-SAMN31437454 contain raw sequencing reads from RNA-seq, whereas raw sequencing reads from target enrichment have been made available under SAMN31439307-SAMN31439402. Additional data and tables have been deposited on Dryad (see Ortiz-Sepulveda et al. (2022); https://datadryad.org/stash/dataset/doi:10.5061/dryad.4j0zpc8gc). Bioinformatic scripts are available via www.github.com/bertvanbocxlaer/target_enrichment_ORF_UCE.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://datadryad.org/stash/dataset/doi:10.5061/dryad.4j0zpc8gc.

BENEFIT-SHARING STATEMENT

Benefits from this research accrue from sharing our data, bioinformatic scripts and results via public databases, as described above. These data contribute to the conservation and sustainable utilization of biological diversity. Additionally, we are committed to international scientific collaboration and cooperation, notably in education and training, and to institutional capacity building.

ORCID

Claudia M. Ortiz-Sepulveda https://orcid.org/0000-0003-0072-719X
Mathieu Genete https://orcid.org/0000-0002-9640-8793
Christian Albrecht https://orcid.org/0000-0002-1490-1825
Xavier Vekemans https://orcid.org/0000-0002-4836-4394
Bert Van Bocxlaer https://orcid.org/0000-0003-2033-326X

REFERENCES

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources*, *20*, 892–905.

Andrews, S. (2010). FastQC: *A quality control tool for high throughput sequence data*. Babraham Institute. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A.

D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477.

Barnosky, A. D., Hadly, E. A., Bascompte, J., Berlow, E. L., Brown, J. H., Fortelius, M., Getz, W. M., Harte, J., Hastings, A., Marquet, P. A., Martinez, N. D., Mooers, A., Roopnarine, P., Vermeij, G., Williams, J. W., Gillespie, R., Kitzes, J., Marshall, C., Matzke, N., … Smith, A. B. (2012). Approaching a state shift in Earth's biosphere. *Nature*, *486*, 52–58.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., & Stadler, P. F. (2013). MITOS: Improved *de novo* Metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, *69*(2), 313–319.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*, 403. https://doi.org/10.1186/1471-2164-1113-1403

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120.

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, *27*(8), 1767–1780.

Borowiec, M. L. (2016). AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, *4*, e1660. https://doi.org/10.7717/peerj.1660

Borowiec, M. L. (2019). Spruceup: Fast and flexible identifications, visualization, and removal of outliers from large multiple sequence alignments. *Journal of Open Source Software*, *4*(42), 1635. https://doi.org/10.21105/joss.01635

Bossert, S., & Danforth, B. N. (2017). On the universality of target-enrichment baits for phylogenomic research. *Methods in Ecology and Evolution*, *9*, 1453–1460.

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources*, *16*, 1059–1068.

Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017). Enriching the ant tree of life: Enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, *8*, 768–776.

Breinholt, J. W., Earl, C., Lemmon, A. R., Lemmon, E. M., Xiao, L., & Kawahara, A. Y. (2018). Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. *Systematic Biology*, *67*(1), 78–93.

Breton, S., Doucet-Beaupré, H., Stewart, D. T., Piontkivska, H., Karmakar, M., Bogan, A. E., Blier, P. U., & Hoeh, W. R. (2009). Comparative mitochondrial genomics of freshwater mussels (Bivalvia: Unionoida) with doubly uniparental inheritance of mtDNA: Gender-specific open reading frames and putative origins of replication. *Genetics*, *183*, 1575–1589.

Breton, S., Stewart, D. T., Shepardson, S., Trdan, R. J., Bogan, A. E., Chapman, E. G., Ruminas, A. J., Piontkivska, H., & Hoeh, W. R. (2011). Novel protein genes in animal mtDNA: A new sex determination system in freshwater mussels (Bivalvia: Unionoida)? *Molecular Biology and Evolution*, *28*(5), 1645–1659.

Buenaventura, E., Lloyd, M. W., Perilla López, J. M., González, V. L., Thomas-Cabianca, A., & Dikow, T. (2021). Protein-encoding ultraconserved elements provide a new phylogenomic perspective of Oestroidea flies (Diptera: Calyptratae). *Systematic Entolomogy*, *46*, 5–27.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). Blast+: Architecture and applications. *BMC Bioinformatics*, *10*, e421. https://doi.org/10.1186/1471-2105-1110-1421

Chan, K. O., Hutter, C. R., Wood, P. L. J., Grismer, L. L., & Brown, R. M. (2020). Larger, unfiltered datasets are more effective at resolving phylogenetic conflict: Introns, exons and UCEs resolve ambiguities in golden-backed frogs (Anura: Ranidae; genus *Hylarana*). *Molecular Phylogenetics and Evolution*, *151*, 106899. https://doi.org/10.1016/j.ympev.2020.106899

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, *15*(5), 538–543.

Cohen, A. S., Stone, J. R., Beuning, K. R. M., Park, L. E., Reinthal, P. N., Dettman, D., Scholz, C. A., Johnson, T. C., King, J. W., Talbot, M. R., Brown, E. T., & Ivory, S. J. (2007). Ecological consequences of early late Pleistocene megadroughts in tropical Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 16422–16427.

Combosh, D. J., Collins, T. M., Glover, E. A., Graf, D. L., Harper, E. M., Healy, J. M., Kawauchi, G. Y., Lemer, S., McIntyre, E., Strong, E. E., Taylor, J. D., Zardus, J. D., Mikkelsen, P. M., Giribet, G., & Bieler, R. (2017). A family-level tree of life for bivalves based on a sanger-sequencing approach. *Molecular Phylogenetics and Evolution*, *107*, 191–208.

Cooney, C. R., Bright, J. A., Capp, E. J. R., Chira, A. M., Hughes, E. C., Moody, C. J. A., Nouri, L. O., Varley, Z. K., & Thomas, G. H. (2017). Mega-evolutionary dynamics of the adaptive radiation of birds. *Nature*, *542*, 344–347.

Cruscuolo, A., & Gribaldo, S. (2010). BMGE (block mapping and gathering with entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Ecology and Evolution*, *10*, 210. https://doi.org/10.1186/1471-2148-1110-1210

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*, giab008. https://doi.org/10.1093/gigascience/giab1008

De Wit, P., & Palumbi, S. R. (2012). Transcriptome-wide polymorphisms of red abalone (*Haliotis rufescens*) reveal patterns of gene flow and local adaptation. *Molecular Ecology*, *21*, 2884–2897.

Dirzo, R., Young, H. S., Galleti, M., Ceballos, G., Isaac, N. J. B., & Collen, B. (2014). Defaunation in the anthropocene. *Science*, *345*, 401–406.

Dunn, C. W., Howison, M., & Zapata, F. (2013). Agalma: An automated phylogenomics workflow. *BMC Bioinformatics*, *14*, e330.

Dutoit, L., Burri, R., Nater, A., Mugal, C. F., & Ellegren, H. (2017). Genomic distribution and estimation of nucleotide diversity in natural populations: Perspectives from the collared flycatcher (*Ficedula albicollis*) genome. *Molecular Ecology Resources*, *17*, 586–597.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Edwards, S. V., Cloutier, A., & Baker, A. J. (2017). Conserved nonexonic elements: A novel class of marker for phylogenomics. *Systematic Biology*, *66*(6), 1028–1044.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., Künstner, A., Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., Uebbing, S., & Wolf, J. B. W. (2012). The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, *491*, 756–760.

Erwin, D. H. (2000). Macroevolution is more than repeated rounds of microevolution. *Evolution & Development*, *2*, 78–84.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*, 2611–2620.

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788.

Faircloth, B. C. (2017). Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods in Ecology and Evolution*, *8*(9), 1103–1112.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*, 717–726.

Froufe, E., Bolotov, I., Aldridge, D. C., Bogan, A. E., Breton, S., Gan, H. M., Kovitvadhi, U., Kovitvadhi, S., Riccardi, N., Secci-Petretto, G., Sousa, R., Teixeira, A., Varandas, S., Zanatta, D., Zieritz, A., Fonseca, M. M., & Lopes-Lima, M. (2020). Mesozoic mitogenome rearrangements and freshwater mussel (Bivalvia: Unionoidea) macroevolution. *Heredity*, *124*, 182–196.

Galtier, N., Gouy, M., & Gautier, C. (1996). SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*, *12*(6), 543–548.

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., Lourenco, J. M., Alves, P. C., Ballenghien, M., Faivre, N., Belkhir, K., Cahais, V., Loire, E., Bernard, A., & Galtier, N. (2013). Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genetics*, *9*(4), e1003457.

Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, *27*(2), 221–224.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321.

Haas, B. J., & Papanicolaou, A. (2018). TransDecoder v. 5.5.0. http://transdecoder.github.io.

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA* [PhD dissertation]. College of Engineering, Pennsylvania State University.

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, *65*(5), 910–924.

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, A., Maes, G. E., Diopere, E., Carvalho, G. R., & Nielsen, E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. *Molecular Ecology Resources*, *11*(Suppl. 1), 123–136.

Hendriks, K. P., Mandáková, T., Hay, N. M., Ly, E., Hooft van Huysduynen, A., Tamrakar, R., Thomas, S. K., Toro-Núñez, O., Pires, J. C., Nikolov, L. A., Koch, M. A., Windham, M. D., Lysak, M. A., Forest, F., Mummenhoff, K., Baker, W. J., Lens, F., & Bailey, C. D. (2021). The best of both worlds: Combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. *Applications in Plant Sciences*, *9*(7), e11438.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522.

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–594.

Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R., & Moussalli, A. (2016). An exon-capture system for the entire class Ophiuroidea. *Molecular Biology and Evolution*, *33*(1), 281–294.

Hutter, C. R., Cobb, K. A., Portik, D. M., Travers, S. L., Wood, P. L. J., & Brown, R. M. (2019). FrogCap: A modular sequence capture probe-set for phylogenomics and population genetics for all frogs, assessed across multiple phylogenetic scales. *Molecular Ecology Resources*, *22*(3), 1100–1119.

Ivory, S. J., Blome, M. W., King, J. W., McGlue, M. M., Cole, J. E., & Cohen, A. S. (2016). Environmental change explains cichlid adaptive radiation at Lake Malawi over the past 1.2 million years. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(42), 11895–11900.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega, N. J., & Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, *4*(7), e1600016.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589.

Karin, B. R., Gamble, T., & Jackman, T. R. (2019). Optimizing phylogenomics with rapidly evolving long exons: Comparison with anchored hybrid enrichment and ultraconserved elements. *Molecular Biology and Evolution*, *37*(3), 904–922.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Korunes, K. L., & Samuk, K. (2021). Pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*, 1359–1368.

Krueger, F. (2019). Trim Galore! Babraham Institute. https://github.com/FelixKrueger/TrimGalore

Kulkarni, S., Wood, H., Lloyd, M., & Hormiga, G. (2020). Spider-specific probe set for ultraconserved elements offers new perspectives on the evolutionary history of spiders (Arachnida, Araneae). *Molecular Ecology Resources*, *20*, 185–203.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Leaché, A. D., & Rannala, B. (2011). The accuracy of species tree estimation under simulation: A comparison of methods. *Systematic Biology*, *60*, 126–137.

Lemmon, A. R., Emme, S., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, *61*, 727–744.

Leroy, T., Rousselle, M., Tilak, M.-K., Caizergues, A. E., Scornavacca, C., Recuerda, M., Fuchs, J., Illera, J. C., De Swardt, D. H., Blanco, G., Thébaud, C., Milá, B., & Nabholz, B. (2021). Island songbirds as windows into evolution in small populations. *Current Biology*, *31*, 1303–1310.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennel, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMTools. *Bioinformatics*, *25*(16), 2078–2079.

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*, 1658–1659.

Li, Y.-L., & Liu, J.-X. (2018). StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. *Molecular Ecology Resources*, *18*, 176–177.

Lopes-Lima, M., Fonseca, M. M., Aldridge, D. C., Bogan, A. E., Ming Gan, H., Ghamizi, M., Sousa, R., Teixeira, A., Varandas, S., Zanatta, D., Zieritz, A., & Froufe, E. (2017). The first Margaritiferidae male (M-type) mitogeonome: Mitochondrial gene order as a potential character for determining higher-order phylogeny within Unionida (Bivalvia). *Journal of Molluscan Studies*, *83*, 249–252.

Lopes-Lima, M., Froufe, E., Do, V. T., Ghamizi, M., Mock, K. E., Kebapçı, Ü., Klishko, O., Kovitvadhi, S., Kovitvadhi, U., Paulo, O. S., Pfeiffer, J. M., III, Raley, M., Riccardi, N., Şereflişan, H., Sousa, R., Teixeira, A., Varandas, S., Wu, X., & Bogan, A. E. (2017). Phylogeny of the most species-rich freshwater bivalve family (Bivalvia: Unionida: Unionidae): Defining modern subfamilies and tribes. *Molecular Phylogenetics and Evolution*, *106*, 174–191.

Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*, 936–939.

Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, *2*, 1940–1955.

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & De Pristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303.

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, *17*, 240–248.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient mehtods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*, 1530–1534.

Moles, J., & Giribet, G. (2021). A polyvalent and universal tool for genomic studies in gastropod molluscs (Heterobranchia). *Molecular Phylogenetics and Evolution*, *155*, 106996. https://doi.org/10.1016/j.ympev.2020.106996

Noé, L., & Kucherov, G. (2005). YASS: Enhancing the sensitivity of DNA similarity search. *Nucleic Acids Research*, *33*, W540–W543.

Ortiz-Sepulveda, C. M., Genete, M., Blassiau, C., Godé, C., Albrecht, C., Vekemans, X., & Van Bocxlaer, B. (2022). Data for: Target enrichment of long open reading frames and ultraconserved elements to link microevolution and macroevolution in non-model organisms. *Dryad*. https://doi.org/10.5061/dryad.4j0zpc8gc

Ortiz-Sepulveda, C. M., Stelbrink, B., Vekemans, X., Albrecht, C., Riedel, F., Todd, J. A., & Van Bocxlaer, B. (2020). Diversification dynamics of freshwater bivalves (Unionidae: Parreysiinae: Coelaturini) indicate historic hydrographic connections throughout the east African rift system. *Molecular Phylogenetics and Evolution*, *148*, 106816. https://doi.org/10.1016/j.ympev.2020.106816

Osbourn, A. E., & Field, B. (2009). Operons. *Cellular and Molecular Life Sciences*, *66*, 3755–3775.

Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., McGuire, J. A., Bowie, R. C., & Moritz, C. (2014). Sequence capture using PCR-generated probes: A cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Molecular Ecology Resources*, *14*, 1000–1010.

Pfeiffer, J. M., Breinholt, J. W., & Page, L. M. (2019). Unioverse: A phylogenomic resource for reconstructing the evolution of freshwater mussels (Bivalvia, Unionoida). *Molecular Phylogenetics and Evolution*, *137*, 114–126.

Portik, D. M., Smith, L. L., & Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (class amphibia, order: Anura). *Molecular Ecology Resources*, *16*, 1069–1083.

Puechmaille, S. J. (2016). The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16*, 608–627.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, *81*, 559–575.

Puritz, J. B., & Lotterhos, K. E. (2018). Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms. *Molecular Ecology Resources*, *18*, 1209–1222.

Quattrini, A. M., Faircloth, B. C., Dueñas, L. F., Bridge, T. C. L., Brugler, M. R., Calixto-Botía, I. F., DeLeo, D. M., Forêt, S., Herrera, S., Lee, S. M., Miller, D. J., Prada, C., Rádis-Baptista, G., Ramírez-Portilla, C., Sánchez, J. A., Rodríguez, E., & McFadden, C. S. (2018). Universal target-enrichment baits for anthozoan (cnidaria) phylogenomics: New approaches to long-standing problems. *Molecular Ecology Resources*, *18*, 281–295.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*, 573–589.

Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., Han, K. L., Harshman, J., Huddleston, C. J., Kingston, S., Marks, B. D., Miglia, K. J., Moore, W. S., Sheldon, F. H., Witt, C. C., Yuri, T., & Braun, E. L. (2017). Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Systematic Biology*, *66*(4), 857–879.

Renaut, S., Guerra, D., Hoeh, W. R., Stewart, D. T., Bogan, A. E., Ghiselli, F., Milani, L., Passamonti, M., & Breton, S. (2018). Genome survey of the freshwater mussel *Venustaconcha ellipsiformis* (Bivalvia: Unionida) using a hybrid *de novo* assembly approach. *Genome Biology and Evolution*, *10*(7), 1637–1646.

Reznick, D. N., & Ricklefs, R. E. (2009). Darwin's bridge between microevolution and macroevolution. *Nature*, *457*, 837–842.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*, 24–26.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J. M., Nabholz, B., Roux, C., Tsagkogeorga, G., Weber, A. A.-T., Weinert, L. A., Belkhir, K., Bierne, N., … Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, *515*, 261–263.

Ronco, F., Matschiner, M., Böhne, A., Boila, A., Büscher, H. H., El Taher, A., Indermaur, A., Malinsky, M., Ricci, V., Kahmen, A., Jentoft, S., & Salzburger, W. (2020). Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature*, *589*(7840), 76–81.

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology*, *14*(12), e2000234.

Schunter, C., Garza, J. C., Macpherson, E., & Pascual, M. (2014). SNP development from RAD-seq data in a nonmodel fish: How many individuals are needed for accurate allele frequency prediction? *Molecular Ecology Resources*, *14*, 157–165.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G. P., Bank, C., Brännström, A., Brelsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., … Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, *15*, 176–192.

Sigwart, J. D., Lindberg, D. R., Chen, C., & Sun, J. (2021). Molluscan phylogenomics requires strategically selected genomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, *376*, 20200161.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, e31. https://doi.org/10.1186/1471-2105-1186-1131

Smit, A. F. A., Hubley, R., & Green, P. (2019). RepeatMasker version 4.0.9. http://repeatmasker.org

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, *26*, 1134–1144.

Sproul, D., Gilbert, N., & Bickmore, W. A. (2005). The role of chromatin structure in regulating the expression of clustered genes. *Nature Reviews Genetics*, *6*, 775–781.

Starrett, J., Derkarabetian, S., Hedin, M., Bryson, R. W., Jr., McCormack, J. E., & Faircloth, B. C. (2017). High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. *Molecular Ecology Resources*, *17*, 812–823.

Streicher, J. W., Miller, E. C., Guerrero, P. C., Correa, C., Ortiz, J. C., Crawford, A. J., Pie, M. R., & Wiens, J. J. (2018). Evaluating methods for phylogenomic analyses, and a new phylogeny for a major frog clade (Hyloidea) based on 2214 loci. *Molecular Phylogenetics and Evolution*, *119*, 128–143.

Sun, J., Mu, H., Ip, J. C. H., Li, R., Xu, T., Accorsi, A., Sánchez Alvarado, A., Ross, E., Lan, Y., Sun, Y., Castro-Vazquez, A., Vega, I. A., Heras, H., Ituarte, S., Van Bocxlaer, B., Hayes, K. A., Cowie, R. H., Zhao, Z., Zhang, Y., & Qiu, J.-W. (2019). Signatures of divergence, invasiveness and terrestralization revealed by four apple snail genomes. *Molecular Biology and Evolution*, *36*(7), 1507–1520.

Tagliacollo, V. A., & Lanfear, R. (2018). Estimating improved partitioning schemes for ultraconserved elements. *Molecular Biology and Evolution*, *35*(7), 1798–1811.

Teasdale, L. C., Köhler, F., Murray, K. D., O'Hara, T., & Moussalli, A. (2016). Identification and qualification of 500 nuclear, single-copy, orthologous genes for the Eupulmonata (Gastropoda) using transcriptome sequencing and exon capture. *Molecular Ecology Resources*, *16*, 1107–1123.

Van Belleghem, S. M., Rastas, P., Papanicolaou, A., Martin, S. H., Arias, C. F., Supple, M. A., Hanly, J. J., Mallet, J., Lewis, J. J., Hines, H. M., Ruiz, M., Salazar, C., Linares, M., Moreira, G. R. P., Jiggins, C. D., Counterman, B. A., McMillan, W. O., & Papa, R. (2017). Complex modular architecture around a simple toolkit of wing pattern genes. *Nature Ecology & Evolution*, *1*, 52. https://doi.org/10.1038/s41559-41016-40052

Van Bocxlaer, B. (2017). Hierarchical structuring of ecological and non-ecological speciation processes of differentiation shaped ongoing gastropod radiation in the Malawi Basin. *Proceedings of the Royal Society B: Biological Sciences*, *284*, 20171494. https://doi.org/10.1098/rspb.2017.1494

Vavouri, T., Walter, K., Gilks, W. R., Lehner, B., & Elgar, G. (2007). Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biology*, *8*, R15. https://doi.org/10.1186/gb-2007-1188-1182-r1115

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548.

Whelan, N. V., Geneva, A. J., & Graf, D. L. (2011). Molecular phylogenetic analysis of tropical freshwater mussels (Mollusca: Bivalvia: Unionoida) resolves the position of *Coelatura* and supports a monophyletic Unionidae. *Molecular Phylogenetics and Evolution*, *61*, 504–514.

Wortley, A. H., Rudall, P. J., Harris, D. J., & Scotland, R. W. (2005). How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. *Systematic Biology*, *54*, 697–709.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J.-Q., & Tian, D. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*, *10*, 47. https://doi.org/10.1186/1471-2164-1110-1147

Zouros, E., Ball, A. O., Saavedra, C., & Freeman, K. R. (1994). An unusual type of mitochondrial DNA inheritance in the blue mussel *Mytilus*. *Proceedings of the National Academy of Sciences of the United States of America*, *91*, 7463–7467.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ortiz-Sepulveda, C. M., Genete, M., Blassiau, C., Godé, C., Albrecht, C., Vekemans, X., & Van Bocxlaer, B. (2023). Target enrichment of long open reading frames and ultraconserved elements to link microevolution and macroevolution in non-model organisms. *Molecular Ecology Resources*, *23*, 659–679. https://doi.org/10.1111/1755-0998.13735